SciGRID_gas: The final IGGIELGNC-2 gas transmission network data set

Release 2.0

J.C. Diettrich,
A. Pluta, J.E. Sandoval, J. Dasenbrock, W. Medjroubi

Jul 01, 2021

CONTENTS

How to cite

J.C. Diettrich, A. Pluta, J.E. Sandoval, J. Dasenbrock, W. Medjroubi
SciGRID_gas: The final IGGIELGNC-2 gas transmission network data set
German Aerospace Center (DLR), Institute for Networked Energy Systems
Germany
doi: 10.5281/zenodo.5017641

Impressum

German Aerospace Center (DLR), Institute for Networked Energy Systems
Carl-von-Ossietzky-Str. 15
26129 Oldenburg
Germany
Tel.: +49 (441) 999 060

LIST OF FIGURES

LIST OF TABLES

Summary

This document describes the final resulting non-OSM data set "IGGIELGNC-2", where all missing values have been estimated using heuristic processes, and was generated by combining the following data sources:

- InternetDaten data set (**INET**) [DPM20e]
- Gas Infrastructure Europe data set (**GIE**) [GasIEurope20]
- Gas Storages Europe data set (**GSE**) [GasSEurope20]
- International Gas Union data set (**IGU**) [IGU20]
- EntsoG-Map data set (**EMAP**) [EntsoG20]
- Long-term Planning and Short-term Optimization data set (**LKD**) [KKS+17]
- Great Britain data set (**GB**) [nationalGrid20]
- Norway data set (**NO**) [Gassco20a]
- Consumers data set for Europe (**CONS**) [San21].

The goal of the SciGRID_gas project is twofold: a) to generate a comprehensive gas transmission network dataset for Europe and b) to develop and supply automated processes to create such data sets for Europe. Gas transmission networks and their data are essential for gas network modelling. The modelling community can derive major characteristics from such networks. Such simulations have a large scope of application, for example, they can be used to perform case scenarios, to model the gas consumption, to minimize leakages and to optimize overall gas distribution strategies. The focus of SciGRID_gas will be on the European transmission gas network, but the principal methods will also be applicable to other geographic regions.

Data required for gas transport models are the gas facilities, such as compressor stations, LNG terminals, pipelines etc. One needs to know their locations, in addition to a large range of attributes, such as pipeline diameter and capacity, compressor capacity, configuration etc. Most of this data is not freely available. However, throughout the SciGRID_gas project it was determined, that data can be grouped into two categories: a) OSM data, and b) non-OSM data. The OSM data consists of geo-referenced facility location data that is stored in the OpenStreetMap (OSM) data base, and is freely available. The OSM data set currently delivers highly accurate topological information on pipelines, however, does rarely contain any required meta information. The Non-OSM data set can fill some of those pipeline data gaps, and can additionally supply information such as pipeline diameter, compressor capacity and more. Part of the SciGRID_gas project is to mine and collate such data, and combine it with the OSM data set. Tools have been designed to fill data gaps and handle copy right issues. This will result in a complete gas network data set.

In this document, the chapter "Introduction" will supply some background information on the SciGRID_gas project. The generated data will be presented in the chapter "Final data set", which gives a brief overview on each set of components and in addition summarizes the raw and estimated attribute values. This data set presented here is slighly different to the previously published data set "IGGIELGNC-3" [DPS+21], as the "IGGIELGNC-2" data set uses roughly 299 European consumer elements, whereas the "IGGIELGNC-3" uses more than 1300 cconsumer elements. Due to the fewer number of consumers, the network could be aggregated, resulting in fewer nodes and pipelines, when compared with the "IGGIELGNC-3" data set.

The appendix contains a glossary, references, graphical results of all heuristic attribute generation processes, location name alterations conventions and finishes with the table of country abbreviation.

CHAPTER

ONE

INTRODUCTION

SciGRID_gas is a three-year project funded by the German Federal Ministry for Economic Affairs and Energy [BMWi20] within the funding of the 6. Energieforschungsprogramm der Bundesregierung [BMWi11].

The goal of SciGRID_gas is to develop methods to generate and provide an open-source gas network data set and corresponding code. Gas transmission network data sets are fundamental for the simulations of the gas transmission within a network. Such simulations have a large scope of application, for example, they can be used to preform case scenarios, to model the gas consumption, to detect leaks and to optimize overall gas distribution strategies. The focus of SciGRID_gas will be the generation of a data set for the European Gas Transmission Network, but the principal methods will also be applicable to other geographic regions.

Both the resulting method code and the derived data will be published free of charge under appropriate open-source licenses in the course of the project. This transparent data policy shall also help new potential actors in gas transmission modelling, which currently do not possess reliable data of the European Gas Transmission Network. It is further planned to create an interface to SciGRID_power [MMK16] or heat transmission networks. Simulations on coupled networks are of major importance to the realization of the German *Energiewende*. They will help to understand mutual influences between energy networks, increase their general performance and minimize possible outages to name just a few applications.

This project was initiated, and is managed and conducted by the German Aerospace Center (DLR), Institute for Networked Energy Systems.

1.1 Project information

- **Project title**: Open Source Reference Model of European Gas Transport Networks for Scientific Studies on Sector Coupling (*Offenes Referenzmodell europäischer Gastransportnetze für wissenschaftliche Untersuchungen zur Sektorkopplung*)
- **Acronym**: SciGRID_gas (Scientific GRID gas)
- **Funding period**: January 2018 - July 2021
- **Funding agency**: Federal Ministry for Economic Affairs and Energy (*Bundesministerium für Wirtschaft und Energie*), Germany
- **Funding code**: Funding Code: 03ET4063
- **Project partner**: German Aerospace Center (DLR), Institute for Networked Energy Systems.

Gefördert durch:

aufgrund eines Beschlusses
des Deutschen Bundestages

1.2 Background

As of today, only limited data of the facilities of the European Gas Transmission Networks is publicly available, even for non-commercial research and related purposes. The lack of such data renders attempts to verify, compare and validate high resolution energy system models, if not impossible. The main reason for such sparse gas facility data is often the unwillingness of transmission system operators (TSOs) to release such commercially sensitive data. Regulations by EU and other lawmakers are forcing the TSOs to release some data. However, such data is sparse and too often not clearly understandable for non-commercial users, such as scientists.

Hence, details of the gas transmission network facilities and their properties are currently only integrated in in-house gas transmission models which are not publicly available. Thus, assumptions, simplifications and the degree of abstraction involved in such models are unknown and often undocumented. However, for scientific research those data sets and assumptions are needed, and consequently the learning curve in the construction of public available network models is rather low. In addition, the commercially sensitivity also hampers any (scientific) discussion on the underlying modelling approaches, procedures and simulation optimization results. At the same time, the outputs of energy system models take an important role in the decision-making process concerning future sustainable technologies and energy strategies. Recent examples of such strategies are the ones under debate and discussion for the Energiewende [BundesregierungDeutschland20] in Germany.

In this framework, the SciGRID_gas project initiated by the German Aerospace Center (DLR), Institute for Networked Energy Systems (Oldenburg, Germany) aims to build an open source model of the European Gas Transmission network. Releasing SciGRID_gas as open-source is an attempt to make reliable data on the gas transmission network available. Appropriate (open) licenses attached to gas transmission network data ensures that established models and their assumptions can be published, discussed and validated in a well-defined and self-consistent manner. In addition to the gas transmission network data, the Python software developed for building the model SciGRID_gas will be published under the GPLv3 license.

The main purpose of the SciGRID_gas project is to provide freely available and well-documented data on the European gas transmission network. Further, with the documentation and the Python code, users should be able to generate the data on their own computers.

The input data itself is based on data available from openstreetmap.org (OSM) under the Open Database License (ODbL) as well as Non-OSM data gathered from different sources, such as Wikipedia pages, fact sheets from TSOs or even newspaper articles.

The main workload of this project is to:

- retrieve the OSM and Non-OSM data sets for the gas infrastructure
- merge all available data sets

- build a gas transmission component data set
- generate missing data using heuristic methods
- document the process and the output.

The first step of the project was to collate a Non-OSM data set by searching the web for metadata that will be useful for the project. This included information, such as pipelines, compressors, LNG terminals, and their attributes, such as diameters, capacities etc. This data set is called the "InternetDaten" data set (INET). The raw data set has been published previously [DPM20e]. Additional data sets, such as the data from "Gas Infrastructure Europe" (GIE), "Gas Storages Europe" (GSE), "International Gas Union" (IGU) and the Norwegian gas transport system operator "Gassco" (NO) are also available. Here all those and other data sets have been merged. In addition, any missing values have been determined using heuristic processes. Other additional data sets will be merged at later stages, and will be made available through the project webpage.

This multi-stage release will allow us to easily and effectively incorporate feedback from potential users during the lifetime of the project. Those releases can be downloaded through the SciGRID_gas webpage with documentation, and can be seen as a snapshot of the current research project state.

The major difference to the previously published IGGIELGNC-3 (see [DPS+21]) is that the consumers were added on a NUTS2 density only, resulting in 294 consumers in the IGGIELGNC-2 when compared with themore than 1300 consumer elements in the IGGIELGNC-3 data set. This resulted also in a smaller number of nodes and pipelines.

Further information on the project can be found on the SciGRID_gas web page: https://www.gas.scigrid.de/pages/imprint.html.

The web page is maintained throughout the project lifetime, and will contain information on:

- General project information
- Contact details
- Presentations
- Bug/data fixes
- Data, code and documentation releases
- Publications.

As part of the SciGRID_gas webpage, one can also sign up to the SciGRID_gas newsletter by sending an email to

news.gas-subscribe@scigrid.de

1.3 Project goal

The overall goals of the SciGRID_gas project are:

- **Data output**: Creation of customisable gas transmission network data sets.
- **Open source**: Any one can download the data, make changes to it, pass it on to others, or even use it in commercial projects, as long as the SciGRID_gas project is mentioned as the original source of the data (CC by).
- **Application**: The outcome of the project can be used for a variety of scientific applications (e.g. sector coupling, entry-exit models etc.).
- **Transparency**: The Python code, the documentation and the data (that can be passed on under copyright licences) is supplied.
- **Extendibility**: Every user can extend the software code to their needs. However, we would encourage users to update and maintain the original git-repository and documentation for others.

- **Feedback**: Through constant data releases, it is hoped that the output data set will improve in quality and quantity by constantly incorporating feedback from the research community.

1.4 Document overview

This is a document describing the **IGGIELGNC-2** data set only. In previous publications, the different input data sets, the SciGRID_gas python data structure, the data merging processes and the framework for generating any missing values has been presented extensivly, and the reader is refered to [DPS+21]. Hence, only the final data set, including data density will be presented, next to some background information that has been placed into the appendix.

1.5 Formatting style

Throughout this document certain editing format styles have been applied, to make it easier for the user to read the document.

Key SciGRID_gas component labels are written in italic, such as *PipeLines*, *Storages* etc.

Component attributes are also written in italic, such as *length_km*, *pressure_bar*.

Function names are written in bold, e.g. **M_CSV.read()**. This also includes build in statistical function, such as **mean** or **median**.

Directory names and file names are surrounded by double quotes, e.g. “StatsMethodsSettings.csv”.

CHAPTER

TWO

FINAL DATA SET

The SciGRID_gas project has the goal of generating a gas transmission network data set for all of Europe. Several individual data sources have been found as part of the project. However, they cannot be used individually, as individual data sets do not contain all the information that is needed for a complete gas transmission network data set. Therefore, several data sets have been combined into a single data set with methods described in previous chapters. After such a process, a significant number of attribute values were still missing in the resulting data set. In [DPS+21] (Chapter Heuristics) described a pathway of how to generate missing attribute values. This resulted in a final gas transmission network data set.

Here the final data set will be described, and differences to a previously published SciGRID_gas data set will be presented as well.

2.1 Combined IGGIELGNC-2 data set

This chapter here will describe the resulting gas transmission network data set, which was constructed by combining the INET, GIE, GSE, IGU, EMAP, LKD, GB, NO and the CON data sets, resulting in the so called IGGIELGNC-2 data set. Each component will be described briefly, mainly focusing on the number of raw versus estimated values. As was described in [DPS+21] two different methods of value estimation have been implemented, a logical/physical based one and a pure statistical based one. It is believed, that summaries should be given for each, so that the reader can get a better understanding of the different methods, and the resulting uncertainties of the approaches.

Here the following terminology in respect of the attribute values will be used:

- "raw": This is referring to the subset of attribute values, that were raw input values.
- "logical/physical": This is referring to the subset of attribute values, that were generated using the logistic/physical methods.
- "statistical": This is referring to the subset of attribute values, that were generated using the pure statistical methods.

2.1.1 PipeSegments

Overall there are 5057 *PipeSegments* elements in the final data set with a length of 206,217 km.

The Table 2.1 depicts the most important attributes that are part of *PipeSegments* elements. The table also presents the number of raw original data, and the number of values that were generated heuristically, including uncertainty values. The table column headings are described below, and will be applicable to all other tables in this chapter here as well:

- "Attribute name": The attribute name.
- "N(R)": The number of raw input values.
- "N(L)": The number of attribute values estimated using a logical/physical base method.

- "N(S)": The number of attribute values estimated using the statistically base method.
- "Ave": Based on the data that is presented in the table, this will be the overall average value of the raw and logical/physical values or the overall average value of the raw and the statistically generated values.
- "Med": Based on the data that is presented in the table, this will be the overall median value of the raw and logical/physical values or the overall median value of the raw and the statistically generated values.
- "U(L)": This is the uncertainty for the logical/physical values.
- "U(S)": This is the uncertainty for the statistical values.
- "Z+(L)": The absolute Z-score (Z+) of the attribute value distributions when comparing the raw distribution with the logical/physical values. An absolute value smaller than two indicates that the distributions are the same.
- "Z+(S)": The absolute Z-score (Z+) of the attribute value distributions when comparing the raw distribution with the statistical values. An absolute value smaller than two indicates that the distributions are the same.
- "P(10)": For a given distribution of values (either raw and logical/physical values or raw and statistical values), the value at the 10 % percentile, informing the user of the spread of the data towards the lower values. (Here no assumptions are being made that the distribution is Gaussian or non-Gaussian, as the determination of the percentile is a simple interpolation of the input values, in respect of the percentile).
- "P(90)": For a given distribution of values (either raw and logical/physical values or raw and statistical values), the value at the 90 % percentile, informing the user of the spread of the data towards the higher values.

Table 2.1: List of attributes of *PipeSegments* elements for the IGGIELGNC-2 data sets, for the raw and logical/physical generated values, with additional statistical properties for each attribute.

Attribute name	N(R)	N(L)	Ave	Med	P(10)	P(90)	U(L)	Z+(L)
diameter_mm	1583	2	871	900	500	1220	139	1.4
is_bothDirection	133	121	0.89	1.00	0.00	1.00	0.10	6.1
max_cap_M_m3_per_d	152	795	35	30	4.0	65	0.32	5.2
max_pressure_bar	919	95	72	70.0	55.0	90.0	1.3	4.3

Table 2.2: List of attributes of *PipeSegments* elements for the IGGIELGNC-2 data sets, for the raw and statistically generated values, with additional statistical properties for each attribute.

Attribute name	N(R)	N(S)	Ave	Med	P(10)	P(90)	U(S)	Z+(S)
diameter_mm	1583	3472	891	900	700	1016	250	4.0
is_H_gas	2449	2608	0.96	1.00	1.00	1.00	0.47	15
max_cap_M_m3_per_d	152	4110	28	27	27	27	22	7.7
max_pressure_bar	919	4043	70	70.0	70.0	70.0	13	1.1

Additional attributes that are not supplied, but were part of the attribute generation process are:

- *pipe_class_EMap*
- *pipe_class_LKD*
- *lat_mean*
- *length_km*
- *long_mean*
- *waterDepth_m.*

In addition, other attributes that were part of the data set, but have been removed prior to release are:

- *exact*
- *num_compressor*
- *operator_name*
- *source*.

is_H_gas

The attribute *is_H_gas* has a data density of almost 50 %, and a high average value **Ave** of 0.96, indicating that a large number of pipelines of the input data set transport high calorific gas. *is_H_gas* is an attribute, for which no relation to any other attribute could be determined. Hence, a constant value of "1" was used to fill all missing attribute values of *is_H_gas*, where an uncertainty of "0.5" was also used for those elements. This approach has been applied to all missing *is_H_gas* attribute values for all components, resulting in very large **Z+**-score value.

max_pressure_bar

The attribute *max_pressure_bar* was generated through two different processes, the logistic onee and the statistical one. The results from the logistic one show a slightly different estimated average value, but a much smaller uncertainty value when compared with the values generated through the statistical process.

max_cap_M_m3_per_d

The attribute *max_cap_M_m3_per_d* is a further attribute, where some of the missing values could be generated with the logical heuristic process. Here again, it can be seen, that the associated uncertainty of the logistic heuristic is smaller when compared with the statistical approach. However, the input data set consists of only 152 raw input values, which is a low portion in respect of the number of attribute values that needed to be generated, hence resulting in a large overall uncertainty. In addition, the range of raw input data ranged from a value of 5 to a value of 200. The **Z+**-score is much larger than 2, indicating that the distributions of values between the raw and the generated data sets are quite different, as most of the missing values were generated using the median value of the raw input data as part of the statistical process.

diameter_mm

This data set contained a larger portion of raw values for the attribute *diameter_mm*, where roughly 30 % of this attribute was supplied as raw values. Here again, a large portion of the missing values were generated by using the median of the raw input values. And it can be seen as well from Table 2.1, that only for two pipe elements this attribute could be generated through the logical heuristic.

Before addressing the other components, information on the distribution of the raw and the estimated values are summarized in Chapter 4.6. This has been carried out through histogram plots. An example of those histogram plots is given in Figure 2.1 for the attribute *max_cap_M_m3_per_d* of the component *PipeSegments*. In addition, the same data is presented on a logarithmic Y-axis scale, so that the smaller bin entries can be seen better (Figure 2.2).

Description of the plots:

- The plot shows in green bars the histogram of the raw input data (left y-axis).
- The red bars indicate the histogram of the estimated values (right y-axis).
- The title contains several items of information:
  - Name of the attribute (excluding the unit)

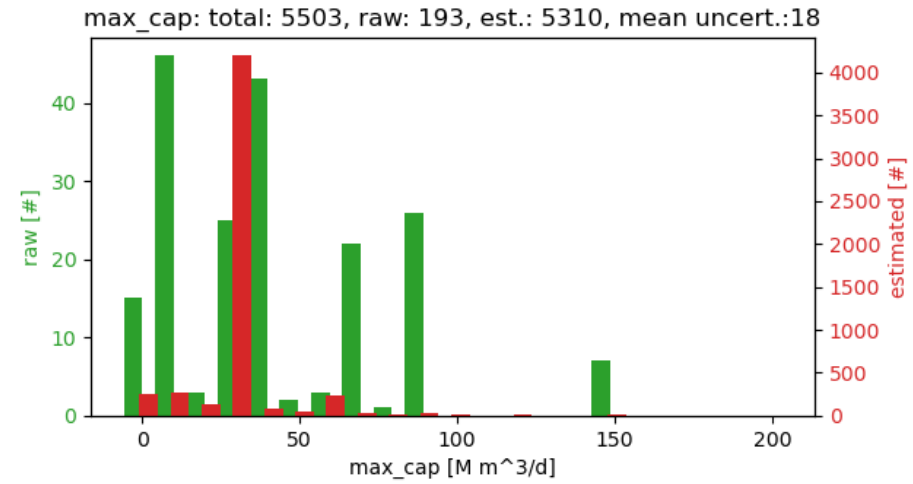


Figure 2.1: Sample plot of the raw and estimated values of the attribute *max_cap_M_m3_per_d* of the component *PipeSegments* (from the IGGIELGNC-3 data set). Green bars are the raw input values, red bars are the histogram of the estimated values. The title and the text below the plot are described in the text below.

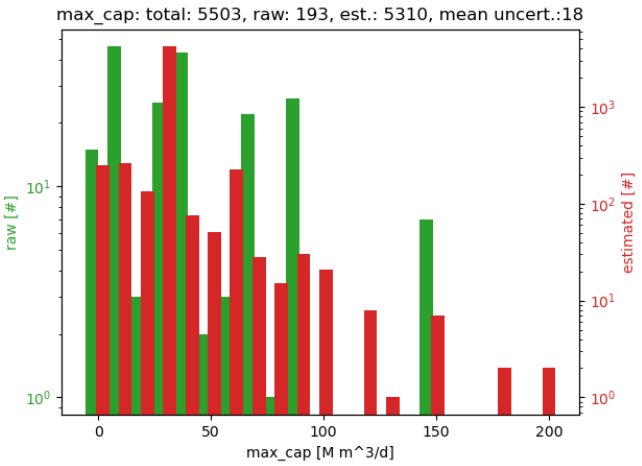


Figure 2.2: Sample plot of the raw and estimated values of the attribute *max_cap_M_m3_per_d* of the component *PipeSegments* on a log Y-axis (from the IGGIELGNC-3 data set). Green bars are the raw input values, red bars are the histogram of the estimated values.

  - Total number of elements of this attribute
  - Number of raw input values
  - The number of generated attribute values
  - The overall mean uncertainty is the last value in the title.
- Below each graph a list of the methods used in generating the missing values is given. Each line is structured as follows:
  - The name of the method
  - In brackets the name(s) of the independent variable
  - The number of attributes that have been generated with these methods
  - In brackets ("mean uncert.") the mean uncertainty of this method for those elements
  - The last line is a summary of the min and maximum raw and estimated values.

The order of the methods listed below the plots does not reflect the order of the application of those methods to generate the missing values. In the automated heuristic attribution generation process, the method with the lowest uncertainties were used before the methods with the higher uncertainties.

With those plots and the additional information in text format below the plots, the user can get an overview of how the missing values were generated, and a summary of their associated uncertainty, hopefully leading to more confidence in the generated data.

2.1.2 Storages

Overall there are 294 *Storages* elements in the final IGGIELGNC-2 data set. The Table 2.3 depicts the most important attributes that are part of the *Storages* elements.

Currently there are no logical/physical heuristics methods implemented for any of the attributes of the component *Storages*. Hence only a single table for the component *Storages* will be presented here, where the raw data will be compared with those values that were generated using the statistical heuristic method.

Table 2.3: List of attributes of *Storages* elements for IGGIELGNC-2 data sets, for the raw and statistically generated values with statistical properties for the most important attributes.

Attribute name	N(R)	N(S)	Ave	Med	P(10)	P(90)	U(S)	Z+(S)
max_cap_store2pipe_M_m3_per_d	185	109	14.0	9.50	2.9	29.0	13.0	4.00
max_cap_pipe2store_M_m3_per_d	175	119	11.0	7.30	2.2	23.0	12.0	3.60
max_cushionGas_M_m3	111	183	893	390	117	1401	1420	3.10
max_power_MW	81	213	14.0	9.10	6.30	16.0	23.0	2.70
is_H_gas	33	261	0.98	1.00	1.00	1.00	0.50	2.40
num_storage_wells	107	187	30.0	19.0	9.00	41.0	36.0	1.70
max_storage_pressure_bar	101	193	131	124	92.0	170	58.0	1.50
max_workingGas_M_m3	194	100	751	304	86	1462	705	0.32
min_storage_pressure_bar	81	213	60.0	60.0	44.3	68.0	25.0	0.19

For the four attributes *max_cap_store2pipe_M_m3_per_d*, *max_cap_pipe2store_M_m3_per_d*, *max_cushionGas_M_m3*, *max_power_MW* and *is_H_gas* one can see that the estimated value significantly changed the distribution of the attribute values, and that the uncertainty in respect of the average value is rather large. Hence here the automated process of generating values did not lead to a satisfactory results.

For the following attributes *num_storage_wells*, *max_storage_pressure_bar*, *max_workingGas_M_m3* and *min_storage_pressure_bar*, the **Z+**-score is smaller than 2, indicating that the estimated distribution is similar to the

input data distribution. However, a closer look at the methods used show that the method median was the dominant method. In addition, the uncertainty values are large as well, when compared with the average of the attributes.

2.1.3 LNGs

Overall there are 32 *LNGs* elements in the final data set. The Table 2.4 depicts all important attributes of the *LNGs* component. Attributes for the component *LNGs* were only given in the data sets INET and GIE, and were explained in previous SciGRID_gas documentations (e.g. [DPM20c]). Here a summary of attribute value distribution for some attributes is given in the Table 2.4.

Table 2.4: List of attributes of *LNGs* elements for the IGGIELGNC-2 data sets, for the raw and statistically generated values, with additional statistical properties for each attribute.

Attribute name	N(R)	N(S)	Ave	Med	P(10)	P(90)	U(S)	Z+(S)
max_cap_store2pipe_M_m3_per_d	30	2	25.7	19.5	10.6	47.7	13.4	2.04
max_workingGas_M_m3	30	2	206	175	49.3	360	64.6	1.94
median_cap_store2pipe_M_m3_per_d	20	12	24.8	19.6	9.63	47.6	3.70	1.46
max_vessel_size_M_m3	22	10	21.7	126	41.2	156	17.4	0.35
GCV_mean_kWh_per_m3	21	11	11.7	11.7	11.6	11.9	0.21	0.01

The **Z+**-score values are larger than 2 for the attributes *max_cap_store2pipe_M_m3_per_d* only. This indicates that the distribution of the raw and the estimated data set are slightly different. One can clearly associate the large difference to the small number of estimated values for the attributes *max_cap_store2pipe_M_m3_per_d*. When comparing the *LNGs* attributes with attributes from other components, one can see, that the uncertainty of the generated data is small, which gives hope to believed, that the automated attribute generation process can be applied for the attributes of the component *LNGs*.

2.1.4 BorderPoints

Overall there are 109 *BorderPoints* elements in the final data set. The overall attribute density for elements of type *BorderPoints* is very high, as for almost all attributes, values were supplied.

Table 2.5: List of attributes of *BorderPoints* elements for the IGGIELGNC-2 data sets, with additional statistical properties for each attribute.

Attribute name	N(R)	N(E)	Ave	Med	P(10)	P(90)	U(E)	Z+
max_cap_to_from_M_m3_per_d	100	9	8.94	0.00	0.00	26.6	9.74	4.33
max_cap_from_to_M_m3_per_d	103	6	26.8	14.7	1.49	64.1	20.1	2.58
GCV_mean_kWh_per_m3	87	22	11.3	11.3	11.1	11.6	0.21	0.22

2.1.5 Compressors

Overall there are 248 *Compressors* elements in the final data set. The Table 2.6 depicts the most important attributes that are part of the *Compressors* component. However, even though the number of compressors was increased through the GB data set, the GB data set did not contain any attribute information, such as *max_cap_M_m3_per_d*, *max_power_MW* or *max_pressure_bar*.

Table 2.6: List of attributes of *Compressors* elements for the IGGIELGNC-2 data sets, with additional statistical properties for each attribute.

Attribute name	N(R)	N(E)	Ave	Med	P(10)	P(90)	U(E)	Z+
is_H_gas	245	3	0.98	1.00	1.00	1.00	0.50	2.02
max_cap_M_m3_per_d	18	230	37.0	37.2	23.5	40.1	23.1	1.65
max_power_MW	36	212	40.1	38.3	38.3	38.3	4.67	1.42
max_pressure_bar	17	231	94.6	94.7	94.7	94.7	6.39	0.08
num_turb	37	211	3.00	3.00	3.00	3.00	0.78	0.12
turbine_fuel_isGas_1	35	213	0.97	1.00	1.00	1.00	0.50	2.96
turbine_fuel_isGas_2	34	214	0.97	1.00	1.00	1.00	0.50	2.97
turbine_fuel_isGas_3	22	226	0.99	1.00	1.00	1.00	0.50	1.48
turbine_power_1_MW	19	229	12.0	11.8	11.8	11.8	4.44	1.67
turbine_power_2_MW	18	230	12.0	11.8	11.8	11.8	4.35	2.04
turbine_power_3_MW	13	235	12.1	12.5	12.5	12.5	4.27	2.91

For the attributes *max_cap_M_m3_per_d*, *max_power_MW* and *max_pressure_bar*, a large number of values needed to be estimated, from an input data set of as little as 17 values. For the attributes *max_pressure_bar*, most of the missing values were generated using the median of the raw input values. For the attribute *max_cap_M_m3_per_d* the missing values were generated using Lasso Linear regressions with other attribute values, hence resulting in varying attribute values, which are similar in distribution to the input values. For the attribute *max_power_MW*, almost all missing values were derived using the method Lasso implemented with the attribute *num_turb* and *turbine_power_4_MW*.

For all other attributes, the **Z+**-score is around 2 or larger than 2, indicating that the distribution of the estimated values is different to the distribution of the raw input data. Here, it needs to be pointed out that all missing values for the attributes *turbine_fuel_isGas_1* to *turbine_fuel_isGas_6* were set to 1 as a blanket rule, as there were no heuristic capabilities of estimating the missing gas type values. For the attribute of the power of the compressors, most missing values were estimated using the median approach, and for turbine power numbers of 4 and larger a value of 0 was applied for all missing values.

2.1.6 Productions

Overall there are 100 *Productions* elements in the final data set. Information for this component mainly came from the EMAP data set, which contained 103 elements throughout Europe, however not supplying any attribute values, such as capacity or start year. The LKD data set is the only other data set that supplied a further 6 elements for Germany, with some information on gas type and maximum production. Here some information for the IGGIELGNC-2 data set will be presented in Table 2.7.

Table 2.7: List of attributes of *Productions* elements for the IGGIELGNC-2 data sets, with additional statistical properties for each attribute.

Attribute name	N(R)	N(E)	Ave	Med	P(10)	P(90)	U(E)	Z+
is_H_gas	6	94	1.00	1.00	1.00	1.00	0.50	N/A
max_supply_M_m3_per_d	5	95	1252	1230	1230	1230	855	0.92

The automated heuristic process achieved lowest fitting uncertainty by using the median of the input data. This was to be expected, as there were only 5 elements with a value for the attribute *max_supply_M_m3_per_d*. Hence all 97 estimated values for *max_supply_M_m3_per_d* have the same value of 1230 $\mathrm{Mm^3 d^{-1}}$, with a large uncertainty of 855 $\mathrm{Mm^3 d^{-1}}$. The uncertainty is very large in respect to the absolute value, which is understandable due to the small number of training values. However, there is no other information in respect of production for those gas production sites.

2.1.7 PowerPlants

Overall there are 316 *PowerPlants* elements in the final data set. Information for this component originated from the INET and the CONS data sets. Here information for the IGGIELGNC-2 data set will be presented in Table 2.8.

Table 2.8: List of attributes of *PowerPlants* elements for the IGGIELGNC-2 data sets, with additional statistical properties for each attribute.

Attribute name	N(R)	N(E)	Ave	Med	P(10)	P(90)	U(E)	Z+
capacity_E_MW	316	0	599	264	26.5	1610	N/A	N/A
capacity_TH_MW	22	8	498	360	44.9	1217	0.00	0.18
is_H_gas	0	316	1.00	1.00	1.00	1.00	0.50	N/A

This data source contained information for all the elements for the attribute *capacity_E_MW*. As no attribute values needed to be estimated for *capacity_E_MW*, the **Z+**-score could not be derived. A further attribute was included for interest only: *capacity_TH_MW* (thermal energy produced for the heat network/processes). Only a small number of power plants came with a value, hence for more than 90 % of power plants, this value was estimated. This does not indicate that all of those power plants are connected to a heat grid. No information was given from the original data sets in that respect. Here, the thermal power estimated is highly correlated to the electric generated installed power *capacity_E_MW*. In addition, it was assumed that all 317 power stations were using the high calorific gas, hence as no raw input data existed, the attribute value *is_H_gas* was set to one.

2.1.8 Consumers

Overall there are 295 NUTS-2 *Consumers* elements in the final data set. Information for this component originated from the INET and the CONS data sets. Here information for the IGGIELGNC-2 data set will be presented in Table 2.9.

Table 2.9: List of attributes of *Consumers* elements for the IGGIELGNC-2 data sets, with additional statistical properties for each attribute.

Attribute name	N(R)	N(E)	Ave	Med	P(10)	P(90)	U(E)	Z+
max_demand_M_m3_per_d	295	0	6.99	6.75	1.59	12.4	0.24	N/A
mean_demand_M_m3_per_d	295	0	2.67	2.40	0.73	4.67	0.24	N/A
median_demand_M_m3_per_d	295	0	2.43	2.23	0.68	4.37	0.24	N/A
min_demand_M_m3_per_d	295	0	1.14	0.90	0.31	2.21	0.24	N/A

This data source contained information for all the elements of all the attributes, as they were generated outside of this project. As no attribute values needed to be estimated for those attributes, the **Z+**-scores were not derived.

2.1.9 Nodes

Overall there are 3539 *Nodes* elements in the final data set. Each original data set contributed some or many *Nodes* elements to the final data set. The nature of the *Nodes* elements is to supply the topological information only. Any latitude and longitude values were derived from the original data set, and any height information was derived using the BING or opentopodata.org web API.

2.1.10 Summary

The IGGIELGNC-2 data set is a further data set created as part of the SciGRID_gas project. For each component the number of elements, and the attribute data density was presented. It was pointed out that for some attributes, other methods of missing value generation need to be found, as the distribution of the estimated attribute values was significantly different to the value distribution of the raw input data. Here further input data and attribute-specific heuristic methods should help in determining "better" missing values. On the other hand, other missing attribute values could easily be generated with the implemented automated attribute generation process.

2.1.11 Resulting map of data set

Below a spatial presentation of the final IGGIELGNC-2 data set is given in Figure 2.3, resulting in a network of 211,814 km in length. In addition, the number of elements for each component is listed in Table 2.10.

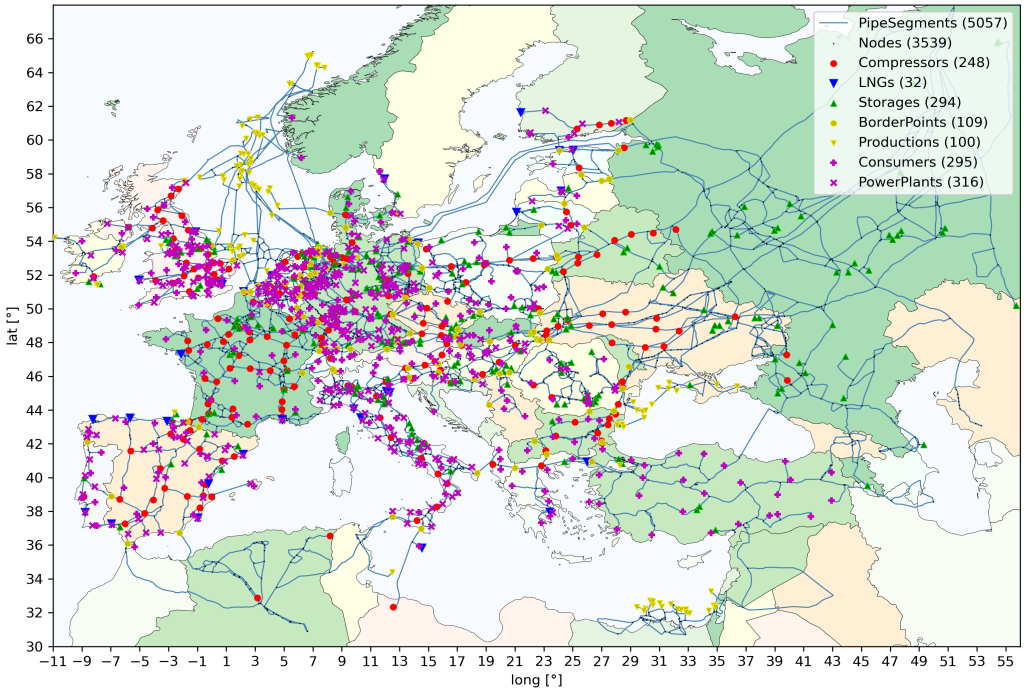


Figure 2.3: Map of the final IGGIELGNC-2 data set.

Table 2.10: List of components with number of elements of the final merged and filled IGGIELGNC-2 network data set.

Component name	Number of elements
BorderPoints	109
Compressors	248
Consumers	295
LNGs	32
Nodes	3539
PipeSegments	5057
PowerPlants	316
Productions	100
Storages	294

CHAPTER

THREE

CONCLUSION

This document describes one of the data sets that are generated as part of the SciGRID_gas project. It starts off with the introduction of the SciGRID_gas project, such as funding, duration and goals. In a subsequent chapter the data structure within the SciGRID_gas project is described, such as components, elements, attributes and attribute values. The third chapter introduced all the different individual data sources: INET, GIE, GSE, IGU, EMap, LKD, GB, NO and CONS data sets. In the next chapter, tools for merging elements are introduced. This is followed by a chapter describing the heuristic generation of any missing attribute value. The final chapter describes briefly the final data set, with its 5057 pipes and more than 240 compressors, where all elements have been connected to a single network, and where all missing attribute values have been estimated using heuristic processes. The final data set is termed "IGGIELGNC-2" data set and spans 211,814 km of transmission pipes over Europe.

CHAPTER

FOUR

APPENDIX

4.1 Glossary

Dataset abbreviations can be found in Table 4.1.

Table 4.1: Dataset abbreviations

Name	Abbreviation	Description
Raw InternetDaten data set	INET	Label/name for the raw InternetDaten data set
Raw Gas Infrastructure Europe data set	GIE	Label/name for the raw Gas Infrastructure Europe data set
Raw Gas Storage Europe data set	GSE	Label/name of the raw Gas Storage Europe data set
Raw Norwegian data set	NO	Label/name for the raw Norwegian data set
Raw Long-term planning and short-term optimization data set	LKD	Label/name for the raw Long-term planning and short-term optimization data set
Raw International Gas Union data set	IGU	Label/name for the raw International Gas Union data set
Raw EntsoG-Map data set	EMAP	Label/name for the raw EntsoG-Map data set
Raw consumer data set	CONS	Label/name for the raw natural gas consumer data set
Merged and filled IGG data set	IGG	Filled data sets, for which the **I**NET, **G**IE and **G**SE data sets were merged
Merged and filled IGGI data set	IGGI	Filled data sets, for which the **I**NET, **G**IE, **G**SE and **I**GU data sets were merged
Merged and filled IGGIN data set	IGGIN	Filled data sets, for which the **I**NET, **G**IE, **G**SE, **I**GU and the **N**O data sets were merged
Merged and filled IGGINL data set	IGGINL	Filled data sets, for which the **I**NET, **G**IE, **G**SE, **I**GU, **N**O and the **L**KD data sets were merged
Merged and filled IGGIELGN data set	IGGIELGN	Filled data sets, for which the **I**NET, **G**IE, **G**SE, **I**GU, **E**MAP, **L**KD, **G**B, and the **N**O data sets were merged
Merged and filled IGGIELGNC-3 data set	IGGIELGNC-3	Filled data sets, for which the **I**NET, **G**IE, **G**SE, **I**GU, **E**MAP, **L**KD, **G**B, and the **N**O data sets were merged, where the CONS data was supplied on a NUTS3 level
Merged and filled IGGIELGNC-2 data set	IGGIELGNC-2	Filled data sets, for which the **I**NET, **G**IE, **G**SE, **I**GU, **E**MAP, **L**KD, **G**B, and the **N**O data sets were merged, where the CONS data was supplied on a NUTS2 level
Merged and filled IGGIELGNC-1 data set	IGGIELGNC-1	Filled data sets, for which the **I**NET, **G**IE, **G**SE, **I**GU, **E**MAP, **L**KD, **G**B, and the **N**O data sets were merged, where the CONS data was supplied on a NUTS1 level

The glossary terms can be found in Table 4.2.

Table 4.2: Glossary

Name	Abbreviation	Description
component		A gas network consists of different components, such as: pipelines, compressors, LNG terminals, storages, entry points and production sites
element		Elements are instances of components. Hence, "10 compressor elements" refers to a data set that contains information for 10 compressor stations
attribute		Gas facilities, such as pipelines or compressors, can be described with a large set of parameters, such as pipeline diameter, or compressor capacity. Those parameters are referred to as attributes
facility		General term used for a gas appliance, such as a single compressor station, or a single LNG terminal
PipeLine		A gas pipeline entity, which has one start and one end point, however, can run via many nodes
PipeSegment		A gas pipeline that has only one start and one end point, but no nodes in-between
LNG	LNG	Liquefied natural gas
CNG	CNG	Compressed natural gas
flow duration curve	FDC	It is the cumulative frequency curve that shows the percentage of time specified flow where equal or exceeded during a given period. The temporal information, when certain events occur, is lost
Energiewende		German term for the change in using primary energies, the move away from coal to renewable energies, such as wind or solar
gas component data set		Raw input data, associated with components of the gas transmission grid
gas network data set		Output data, a coherent network of gas transmission components
OSM	OSM	Data that is available from openstreetmap.org
non-OSM	Non-OSM	Data that is not part of the OSM data set
gas type		There are two types of gas: High (H) and Low (L) calorific gas
mean absolute error	MAE	mean difference between input values and estimated values
data density		The ratio of the number of usable (not missing) attribute values over number elements of the component, in units of [%]
Transmission System Operator	TSO	An entity entrusted with the transportation of natural gas/electricity, as defined by the European Union
gas transmission network		This describes the physical gas transmission grid, however, it excludes any facilities/components that would be part of a distribution network and their facilities
gas component data set		The term "gas component data set" is used for raw data sets of gas network facilities. However, not all elements (e.g. compressors) need to be connected to pipelines, where the emphasis is on the term **component**
gas network data set		A "gas component data set" can be converted into a "gas network data set", by connecting all non-pipeline elements to nodes and all nodes are connected to pipelines. Hence, the emphasis here is on the term **network**
Nomenclature des unités territoriales statistiques	NUTS	Geographical system dividing the Eurpean Union into regions of similar size in respect of number of inhabitatnts

4.2 Unit conversions

Table 4.3: Unit conversions

From Unit	To Unit	MultiVal
LNG Mt	LNG Mm^3	2.47
gas $\mathrm{tm}^3\mathrm{h}^{-1}$	gas $\mathrm{Mm}^3\mathrm{d}^{-1}$	24/1000
LNG Mm^3	gas Mm^3	584
LNG t	gas Mm^3	1442.48
GWh (H)	gas Mm^3	0.0879757777
GWh (L)	gas Mm^3	0.1023541453

For some elements of some components, the calorific value was given through the references. Hence during the conversion process from GWh to M m^3, the elements calorific value was used, however, whereever the element specific calorific value was not known, the default values from Table 4.3 was used in dependence of the gas type of the element. If no gas type was known, then high calorific gas is assumed.

4.3 Attribute *exact*

Each element of type *Nodes* has an attribute *exact*. With this, the SciGRID_gas project is trying to let the user know, how well the actual location of the *Nodes* elements are known. The actual location (latitude-longitude pair) can be spot on (verifiable through satellite imagery) or can be unknown by 10's or 100's of km, where city names or country names are known only. Here the attribute value for *exact* is being given, ranging from "1" to "5" as listed in Table 4.4 below.

Table 4.4: Unit conversions

Exact value	Description	Uncertainty [km]
1	The exact location of this node is known, as one was able to verify the facility through satellite data.	0
2	Here the lat/long is not known exactly. However, one assumes that the location is within a small region (e.g. Krummhörn). Hence, not being much larger than 10 km	10
3	Here so little is known about the exact location, and one only knows that the location is within a large region (e.g. Hamburg). Hence, the actual location could be out by 10 km or more but less than 100 km	100
4	Here so little is known about the exact location, and one only knows that the location is within a state (e.g. Niedersachsen). Hence, the actual location could be out by 100 km or more but less than 1000 km	1000
5	Here so little is known about the exact location, and one only knows that the location is within a country (e.g. Ukraine). Hence, the actual location could be out by 1000 km or more.	> 1000

4.4 Location name alterations

Location names should be changed into the 26 letters used in the English language.

For names from the individual countries please follow the suggested approach:

- Germany/Austria: *Umlaute* to be replaced with the letter followed by an 'e', e.g.: ü = ue.
- France/Belgium: Omit accent de gues and accent de graphs, e.g.: ó = o.
- Sweden: Please change the last three letters of the Swedish alphabet and replace e.g.: ä = a.
- Poland: Please change any letter that cannot be found in the English alphabet, knowing that for some letters that one can only use a single letter instead of the three different letters used in the Polish alphabet, e.g.: z = z.
- Spain/Portugal: Please change any letter that cannot be found in the English alphabet, e.g.: ñ = n.
- Greece: Please do not use Greek letters. Please try to write the Greek words with Latin letters.
- Denmark: Please change any letter that contains non-English letters, e.g.: "å" with "aa".
- Slovakia, Czech Republic, Hungary, Rumania, Latvia, Lithuania, Estonia, Bulgaria, Slovenia, Croatia: PLEASE use your common sense, based on the examples from the other countries above.

4.5 Country name abbreviations

For convenience we provide a short list of names and two-digit codes (see Table 4.5) for the probably most important countries associated with the European Transmission Grid.

Table 4.5: Country codes

Country name	Country code	Country name	Country code
Albania	AL	Kosovo	XK
Armenia	AM	Latvia	LV
Austria	AT	Liechtenstein	LI
Azerbaijan	AZ	Lithuania	LT
Belarus	BY	Luxembourg	LU
Belgium	BE	Malta	MT
Bosnia and Herzegovina	BA	Moldova	MD
Bulgaria	BG	Montenegro	ME
Croatia	HR	Netherlands	NL
Cyprus	CY	Norway	NO
Czech	CZ	Poland	PL
Denmark	DK	Portugal	PT
Estonia	EE	Romania	RO
Finland	FI	Serbia	RS
France	FR	Slovakia	SK
Georgia	GE	Slovenia	SI
Germany	DE	Spain	ES
Greece	GR	Sweden	SE
Hungary	HU	Switzerland	CH
Iceland	IS	Turkey	TR
Ireland and Northern Ireland	IE	Belarus	UA
Italy	IT	Great Britain	GB
Russia Federation	RU	Europe	EU
Ukraine	UA		

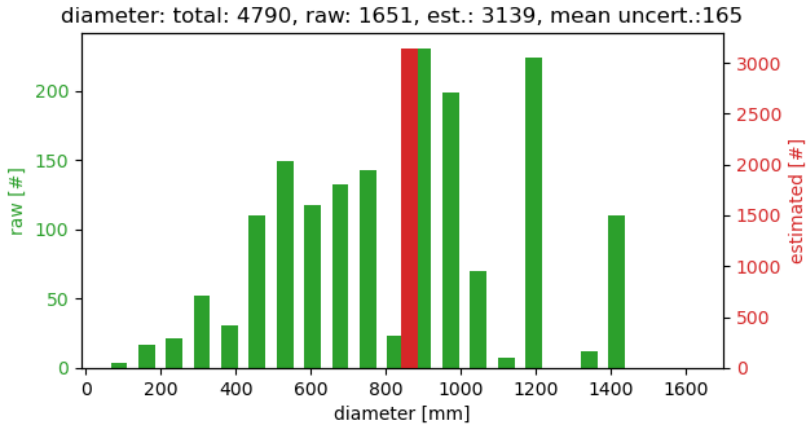
4.6 Heuristic histogram plots of the IGGIELGNC-2 data set

Below, for each filled attribute two histogram plots will be presented. The first plot for each attribute will be the histogram with normal Y-axis scaling, whereas the second plot will depict the histogram on a log Y-axis scale. Each of those plots will show in green bars the histogram of the raw input data (left Y-axis), and with red bars the histogram of the estimated values (right Y-axis). The title contains the name of the attribute, the total number of elements of this attribute, the number of raw input values, the total sum of generated attribute values and the overall mean uncertainty of the attribute values. In addition, below each graph with the linear scale, a list of methods used is given. Each method name is followed by the name of the independent variable or variables, supplied in brackets. This is followed by the number of attribute values that were generated with the methods. The values in the last bracket in each line give the mean uncertainty for those attribute values generated, excluding the raw input data. Further information is given in the last line, where the minimum and maximum values of the raw input data ("min(raw)" and "max(raw)"), and the minimum and maximum of the estimated data ("min(raw)" and "max(raw)") are presented.

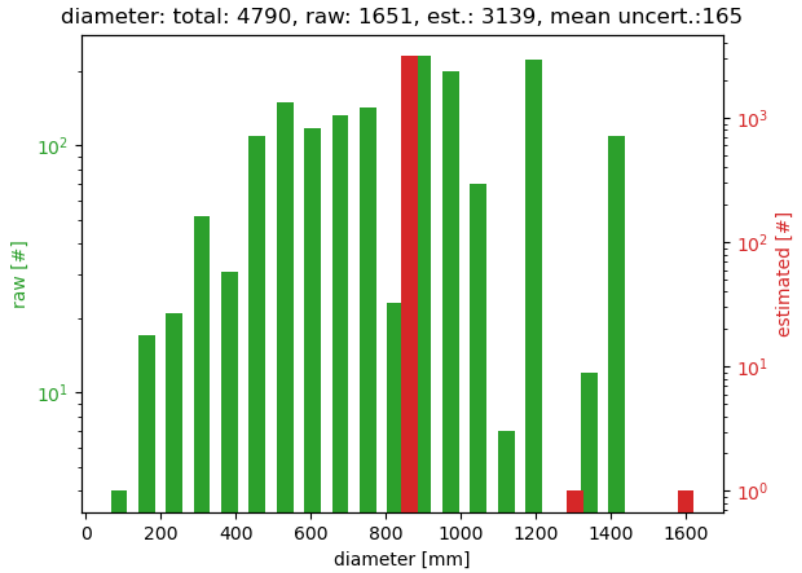
4.6.1 *PipeSegments*

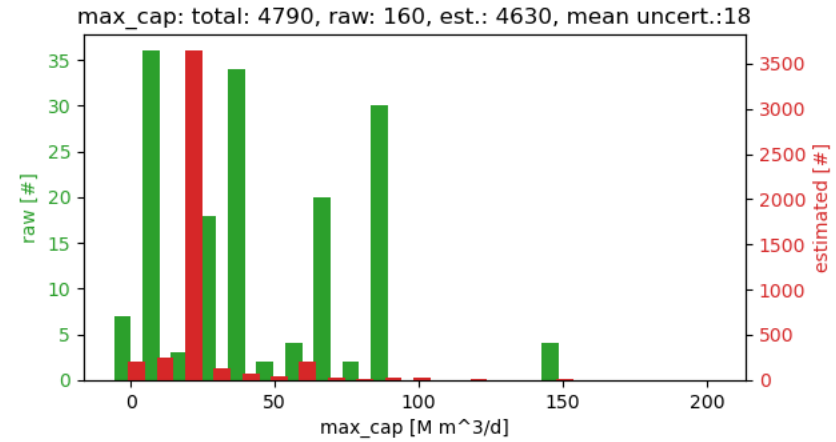
Below are the heuristic histogram plots of the component *PipeSegments* for the attributes:

- *diameter_mm*
- *max_cap_M_m3_per_d*
- *max_pressure_bar*.

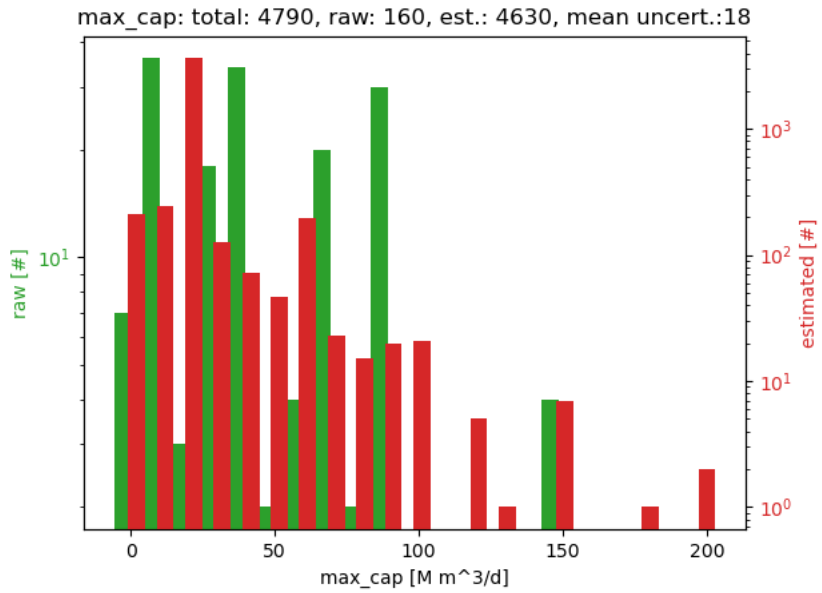


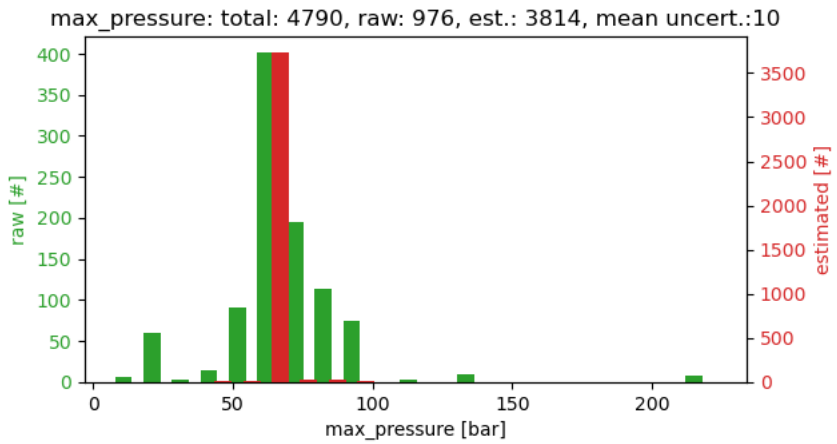
Median(diameter_mm): 3136 (mean uncert.: 251)
M_Heuristics.adjust_AttribValues(): 1 (mean uncert.: 500)
log_M_Heuristics.PipeSegment_diameter(): 2 (mean uncert.: 137)
min(raw):125.0, max(raw):1422.4, min(est):863.6, max(est):1600.0



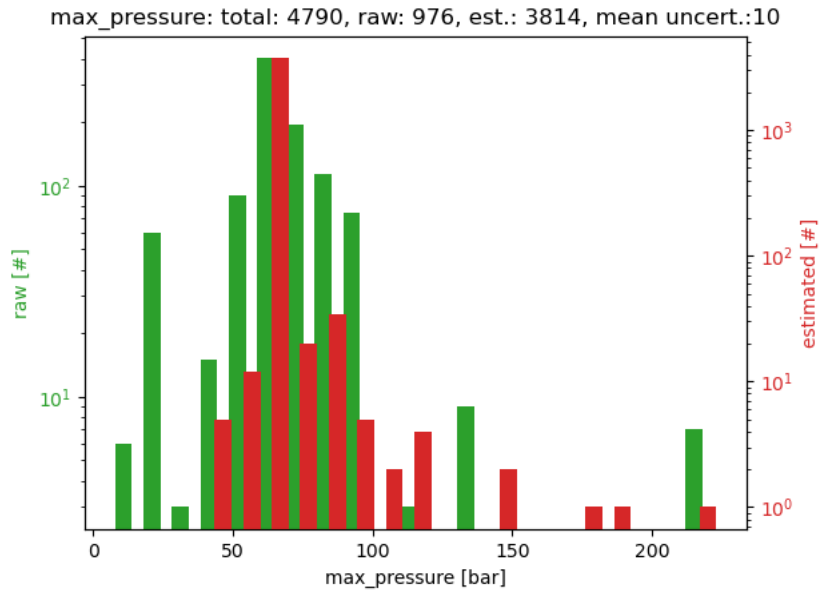


Median(max_cap_M_m3_per_d): 3518 (mean uncert.: 23)
M_Heuristics.app_Phys_Heur_Move_Capacity(): 237 (mean uncert.: 0)
log_M_Heuristics.PipeSegment_capacity(): 350 (mean uncert.: 1)
M_Heuristics.adjust_AttribValues(): 24 (mean uncert.: 50)
log_raw: 499 (mean uncert.: 0)
Squared(Compressors_max_power_MW, Storages_max_cushionGas_M_m3): 2 (mean uncert.: 1)
min(raw):6.3, max(raw):150.68, min(est):2.0, max(est):200.0





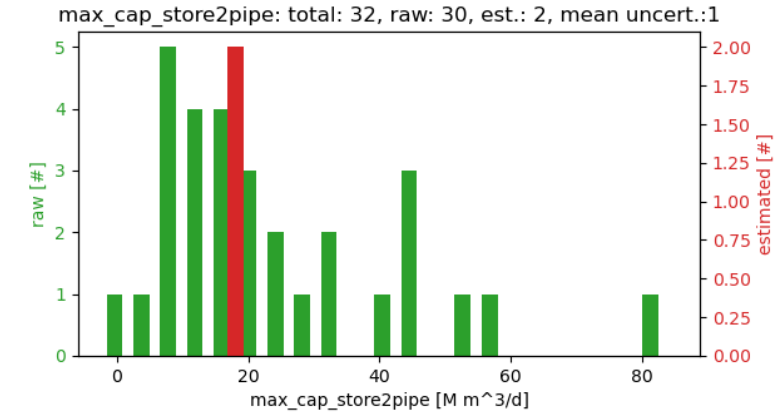
Median(max_pressure_bar): 3712 (mean uncert.: 13)
log_M_Heuristics.PipeSegment_pressure(): 97 (mean uncert.: 1)
M_Heuristics.adjust_AttribValues(): 1 (mean uncert.: 100)
Squared(Compressors_max_power_MW, Storages_max_power_MW): 3 (mean uncert.: 1)
Lasso(Compressors_max_power_MW, Storages_max_cushionGas_M_m3): 1 (mean uncert.: 1)
min(raw):16.0, max(raw):220.0, min(est):44.44, max(est):220.0



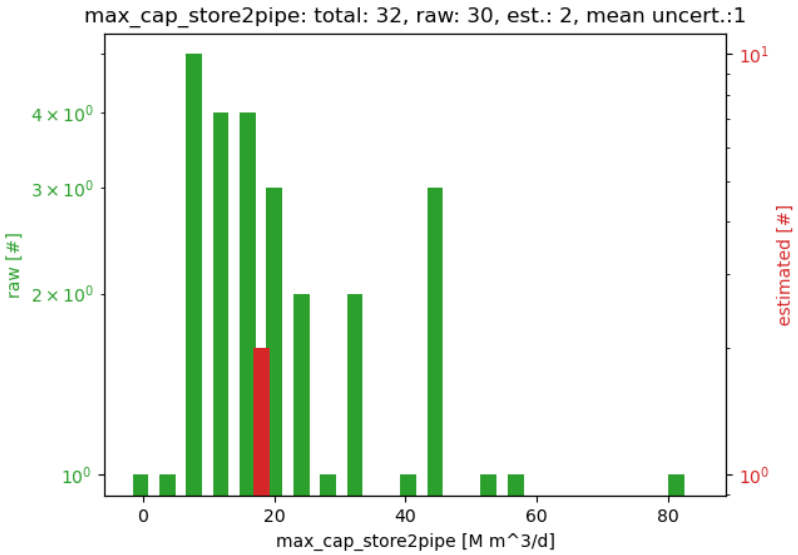
4.6.2 *LNGs*

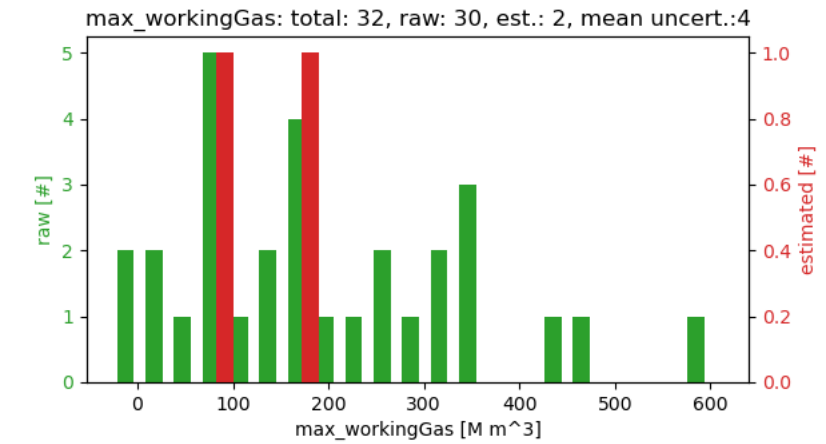
Below are the heuristic histogram plots of the component *LNGs* for the attributes:

- *max_cap_store2pipe_M_m3_per_d*
- *max_workingGas_M_m3*
- *median_cap_store2pipe_M_m3_per_d.*

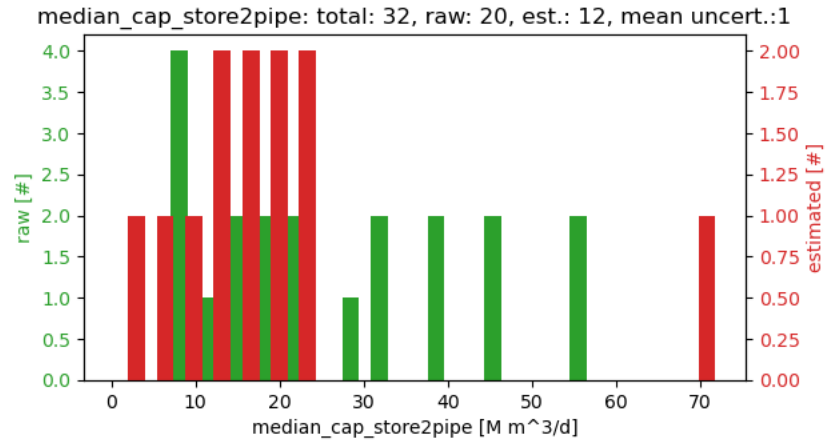


Median(max_cap_store2pipe_M_m3_per_d): 2 (mean uncert.: 13)
min(raw):1.64, max(raw):83.35, min(est):19.38, max(est):19.38

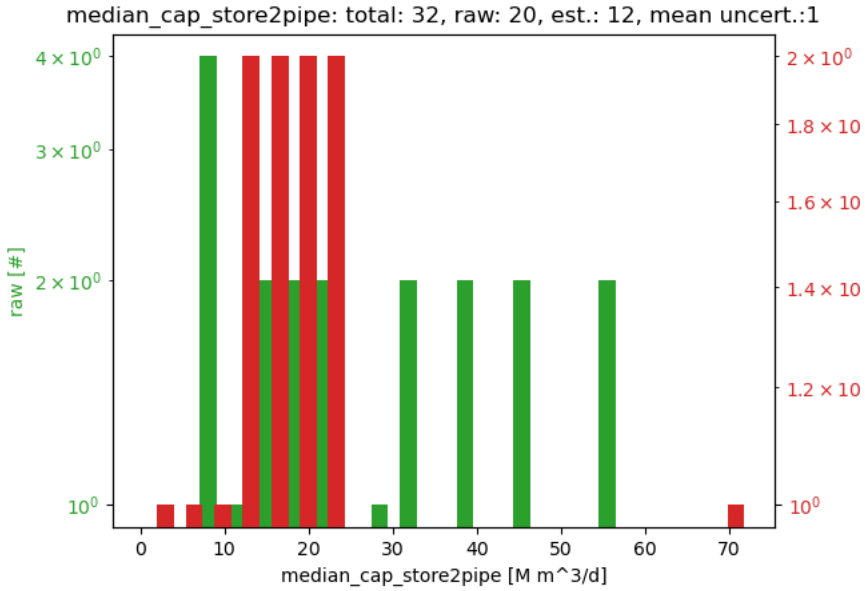




Median(max_workingGas_M_m3): 1 (mean uncert.: 121)
Squared(max_cap_store2pipe_M_m3_per_d, Pipe_max_cap_M_m3_per_d): 1 (mean uncert.: 8)
min(raw):2.32, max(raw):599.42, min(est):104.47, max(est):175.2



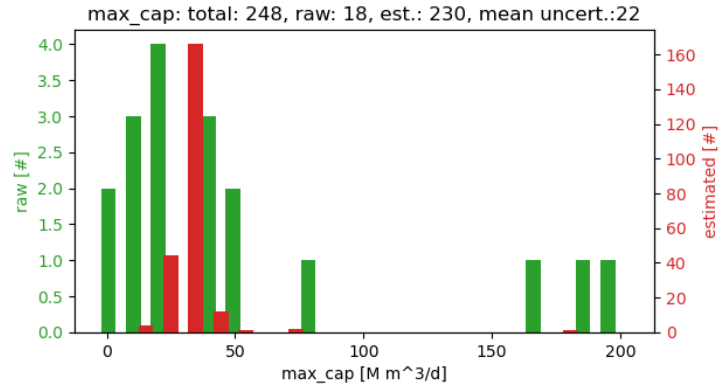
Lasso(max_cap_store2pipe_M_m3_per_d, max_workingGas_M_m3): 8 (mean uncert.: 2)
Median(median_cap_store2pipe_M_m3_per_d): 2 (mean uncert.: 13)
Lasso(max_cap_store2pipe_M_m3_per_d): 2 (mean uncert.: 2)
min(raw):8.44, max(raw):57.18, min(est):3.01, max(est):70.8



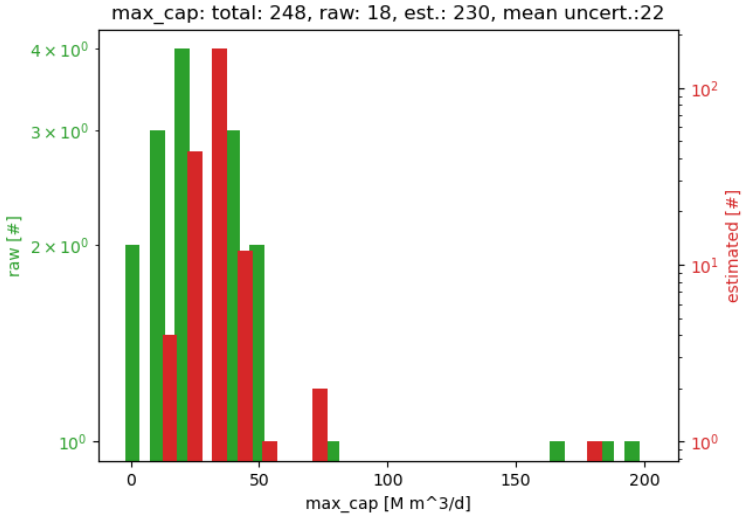
4.6.3 *Compressors*

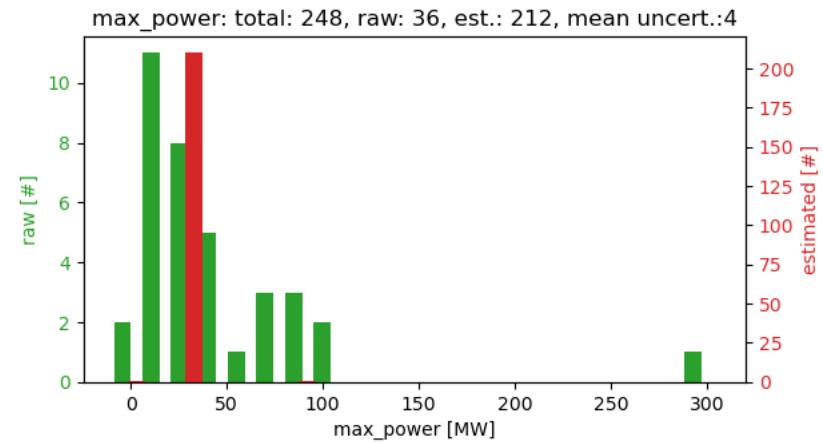
Below are the heuristic histogram plots of the component *Compressors* for the attributes:

- *max_cap_M_m3_per_d*
- *max_power_MW*
- *max_pressure_bar*
- *num_turb*
- *turbine_power_1_MW*
- *turbine_power_2_MW*
- *turbine_power_3_MW*
- *turbine_power_4_MW*.

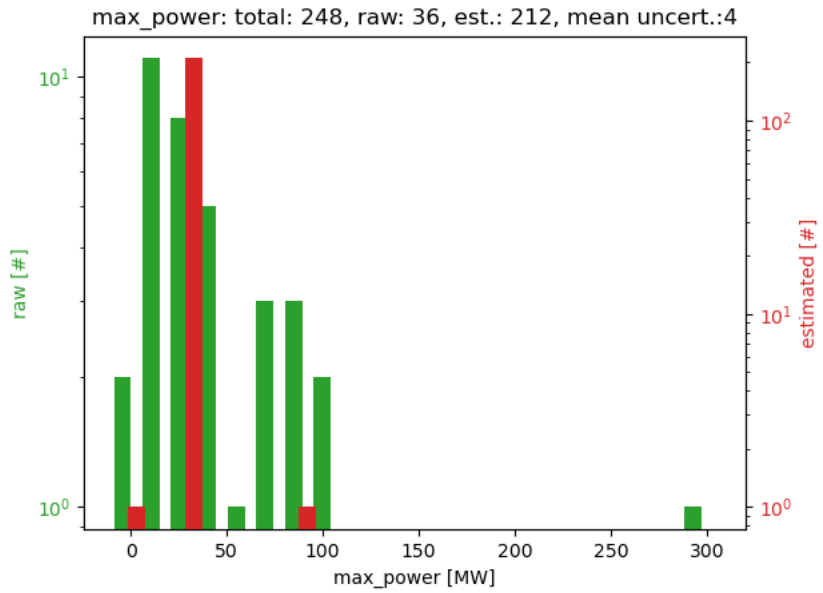


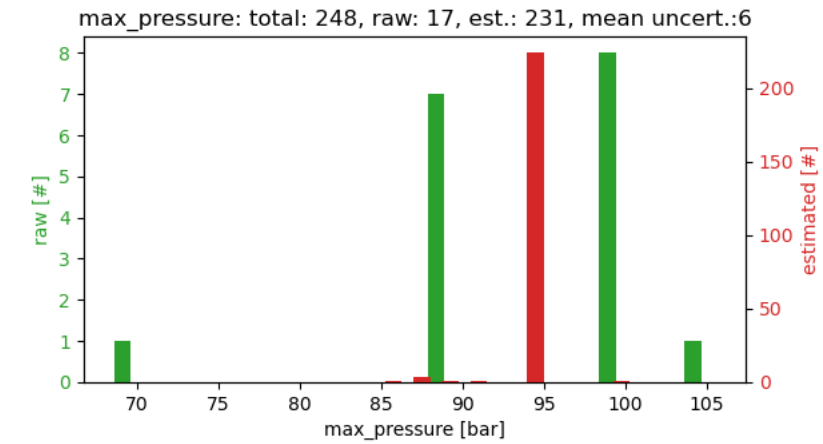
Lasso(max_power_MW, Pipe_diameter_mm): 110 (mean uncert.: 6)
Median(max_cap_M_m3_per_d): 117 (mean uncert.: 40)
Squared(max_power_MW, Pipe_max_cap_M_m3_per_d): 3 (mean uncert.: 7)
min(raw):5.52, max(raw):199.99, min(est):11.94, max(est):180.06



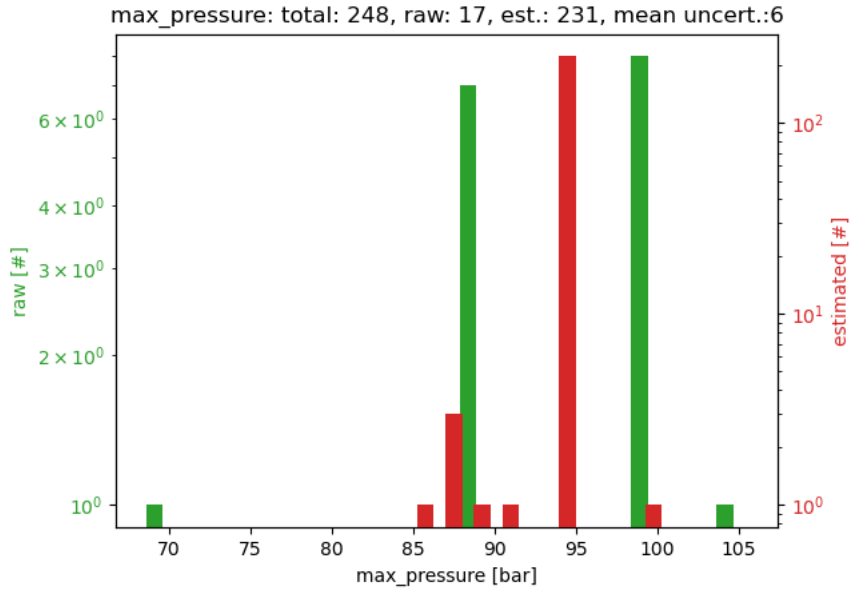


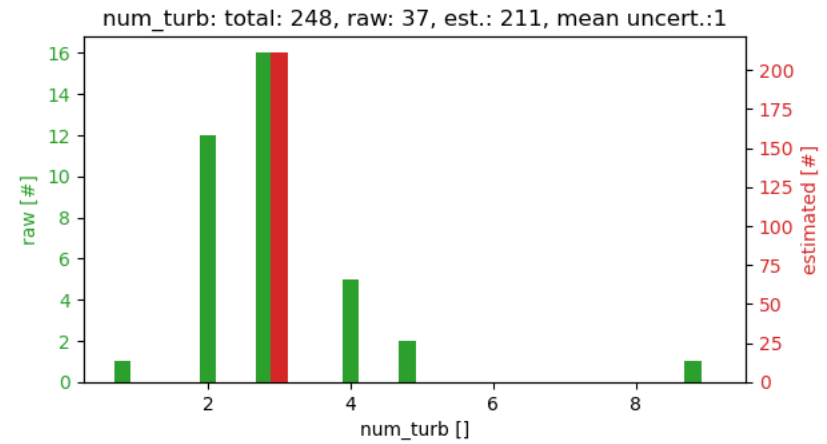
Lasso(num_turb, turbine_power_4_MW): 208 (mean uncert.: 5)
M_Internet.set_max_power_MW: 4 (mean uncert.: 0)
min(raw):5.5, max(raw):300.0, min(est):3.0, max(est):85.6



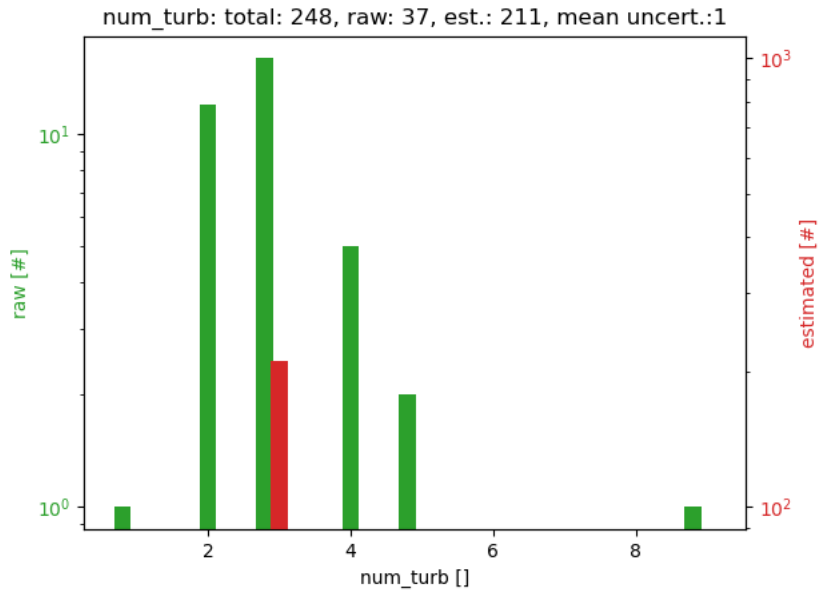


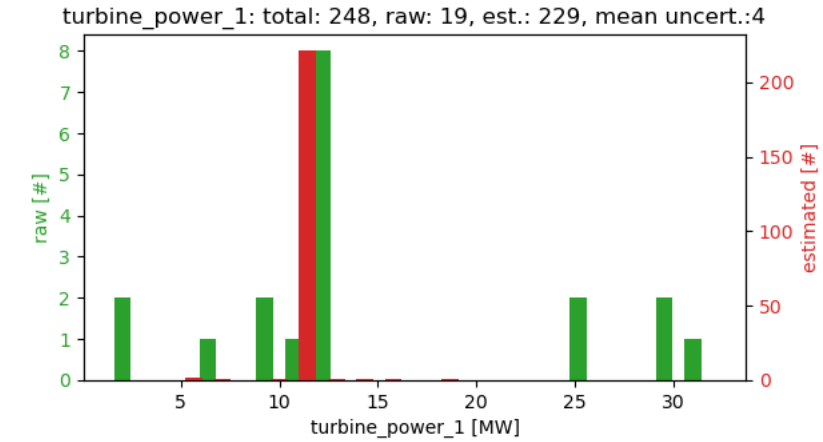
Median(max_pressure_bar): 224 (mean uncert.: 6)
Lasso(turbine_power_2_MW, Pipe_max_cap_M_m3_per_d): 3 (mean uncert.: 4)
Lasso(turbine_power_3_MW, Pipe_max_cap_M_m3_per_d): 4 (mean uncert.: 3)
min(raw):70.0, max(raw):105.0, min(est):86.19, max(est):100.59



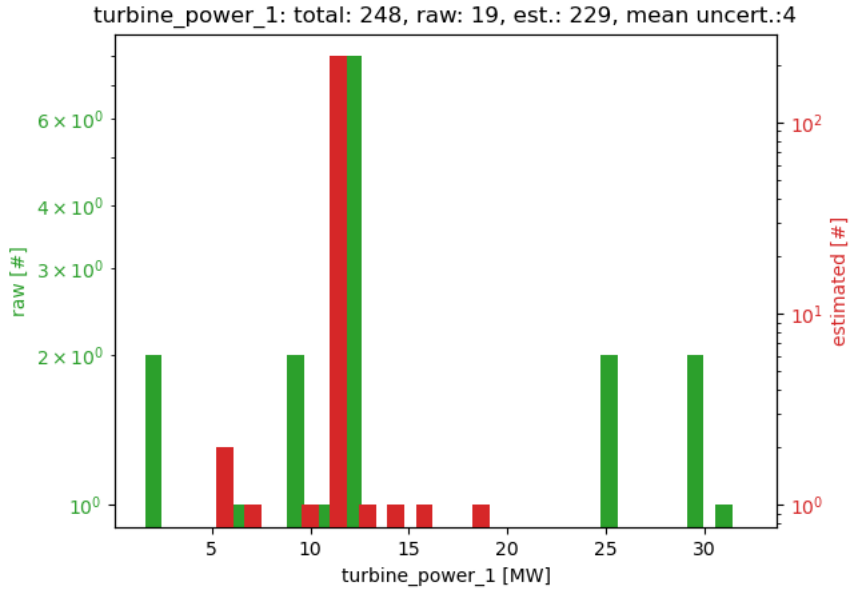


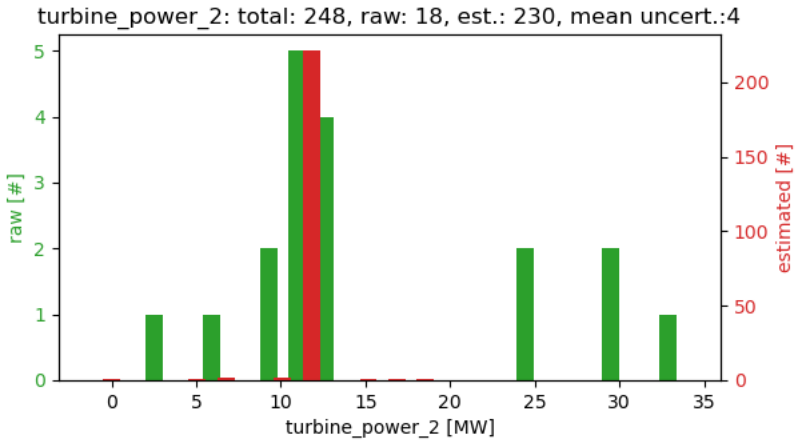
Median(num_turb): 211 (mean uncert.: 1)
min(raw):1, max(raw):9, min(est):3, max(est):3



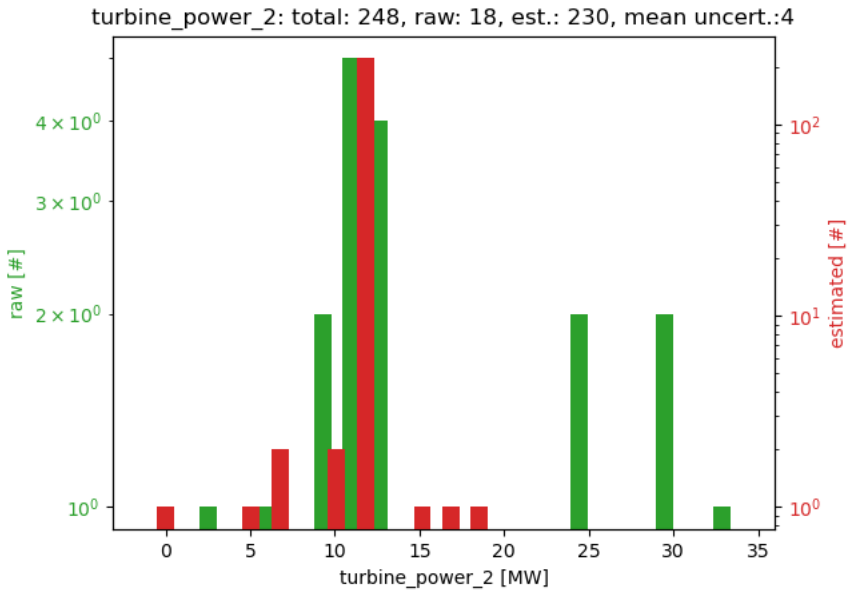


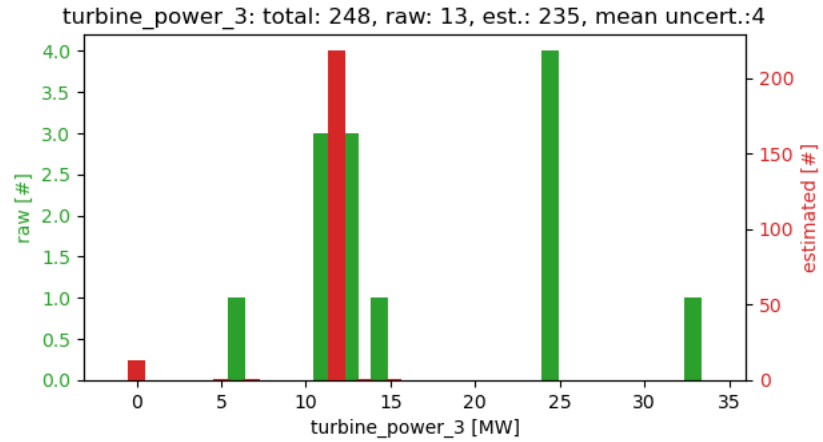
Median(turbine_power_1_MW): 212 (mean uncert.: 4)
M_Internet.set_turbine_power: 17 (mean uncert.: 6)
min(raw):2.75, max(raw):31.7, min(est):5.0, max(est):19.3



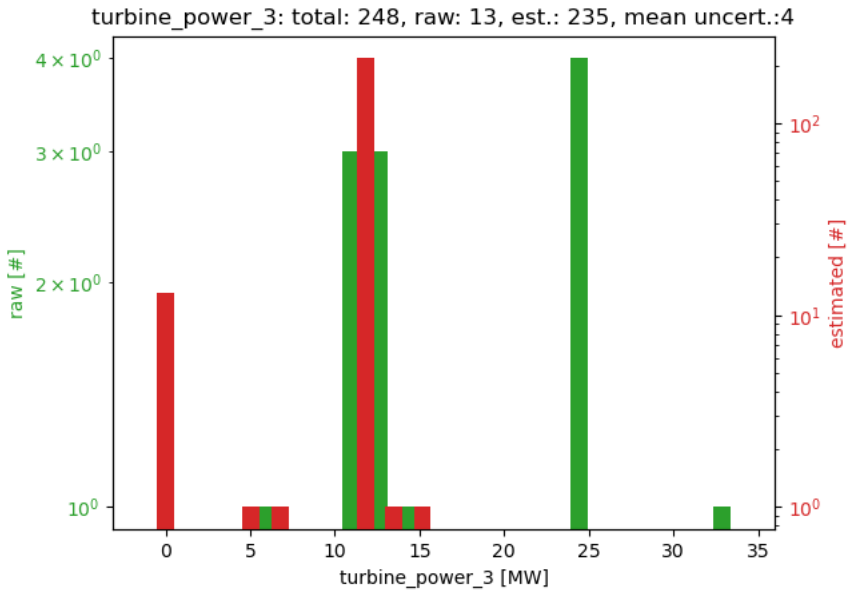


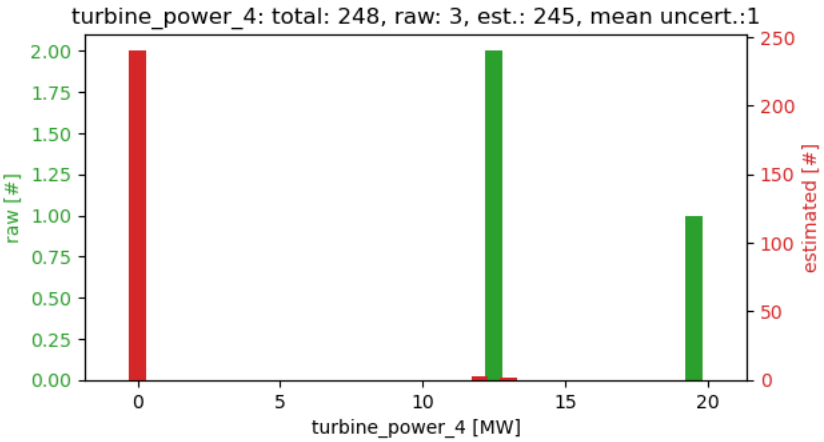
Median(turbine_power_2_MW): 212 (mean uncert.: 4)
M_Internet.set_turbine_power: 17 (mean uncert.: 6)
make_Attrib(const): 1 (mean uncert.: 0)
min(raw):2.75, max(raw):33.7, min(est):0, max(est):19.3



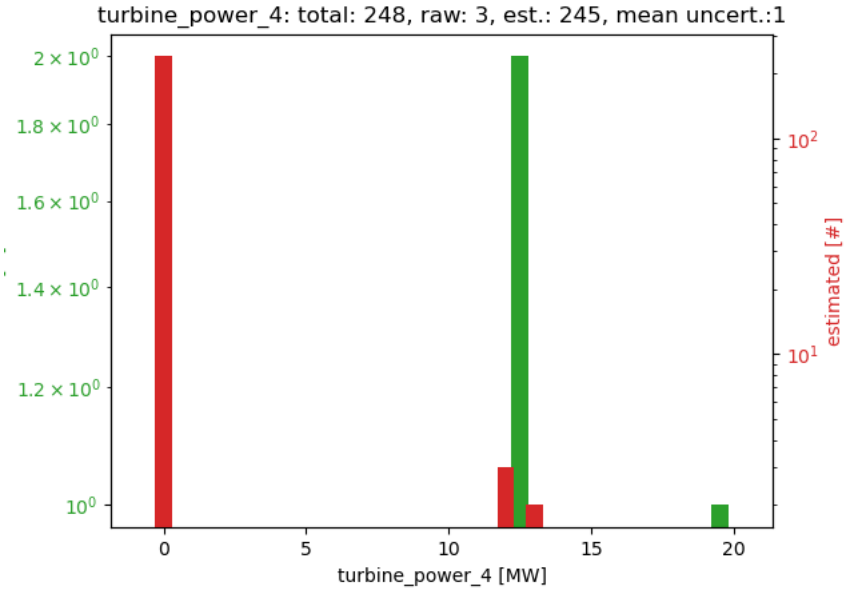


Median(turbine_power_3_MW): 212 (mean uncert.: 4)
make_Attrib(const): 13 (mean uncert.: 0)
M_Internet.set_turbine_power: 9 (mean uncert.: 5)
copied from turbine_power_2_MW: 1 (mean uncert.: 0)
min(raw):7.0, max(raw):33.7, min(est):0, max(est):15.0





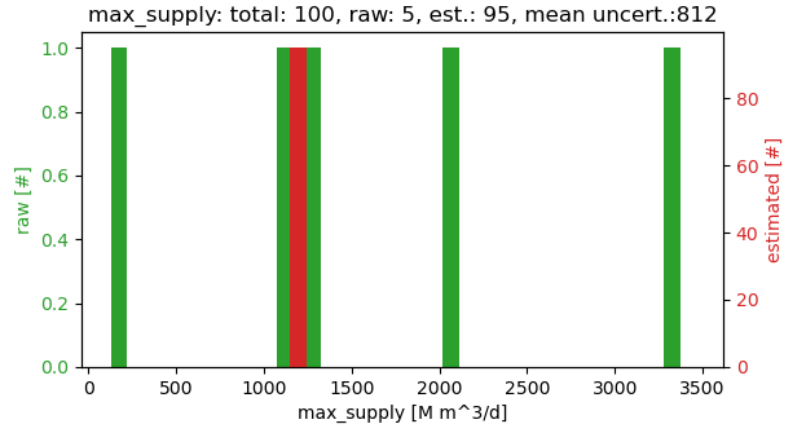
make_Attrib(const): 240 (mean uncert.: 0)
M_Internet.set_turbine_power: 4 (mean uncert.: 6)
copied from turbine_power_3_MW: 1 (mean uncert.: 0)
min(raw):12.5, max(raw):20.0, min(est):0, max(est):13.0



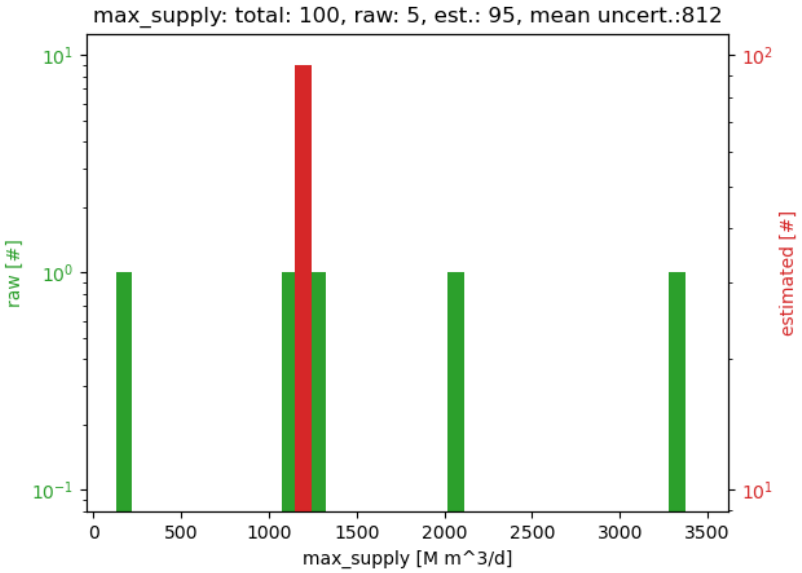
4.6.4 *Productions*

Below are the heuristic histogram plots of the component *Productions* for the attributes:

- *max_supply_M_m3_per_d*.



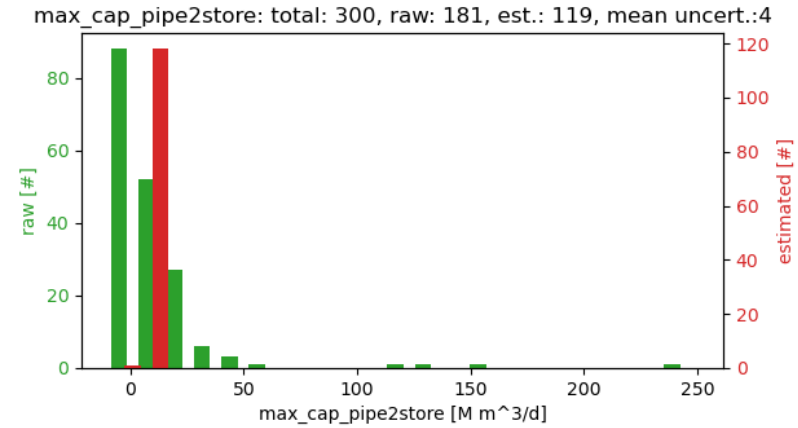
Median(max_supply_M_m3_per_d): 95 (mean uncert.: 855)
min(raw):253.66, max(raw):3403.59, min(est):1230.41, max(est):1230.41



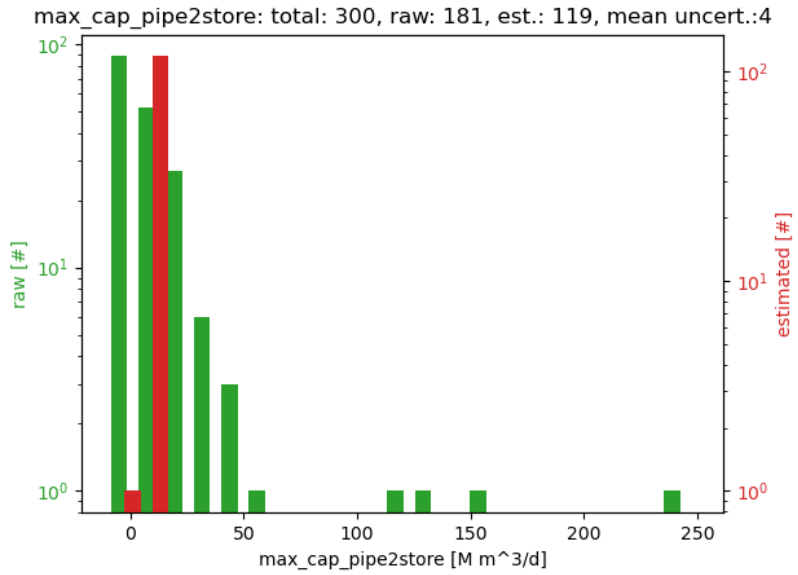
4.6.5 *Storages*

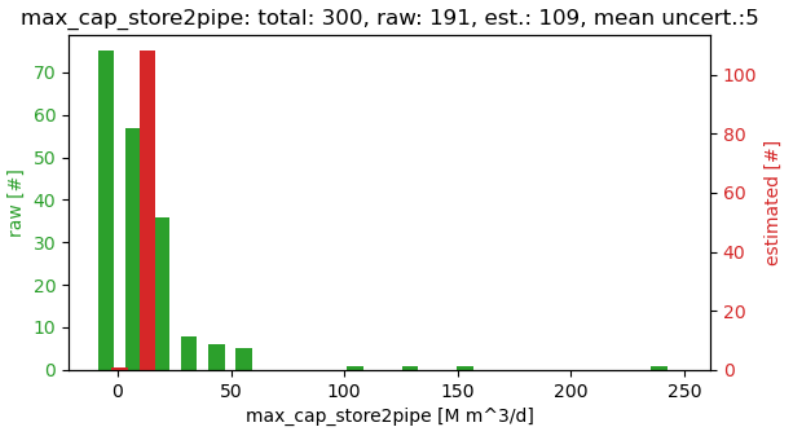
Below are the heuristic histogram plots of the component *Storages* for the attributes:

- *max_cap_pipe2store_M_m3_per_d*
- *max_cap_store2pipe_M_m3_per_d*
- *max_cushionGas_M_m3*
- *max_power_MW*
- *max_storage_pressure_bar*
- *max_workingGas_M_m3*
- *min_storage_pressure_bar*
- *num_storage_wells*.

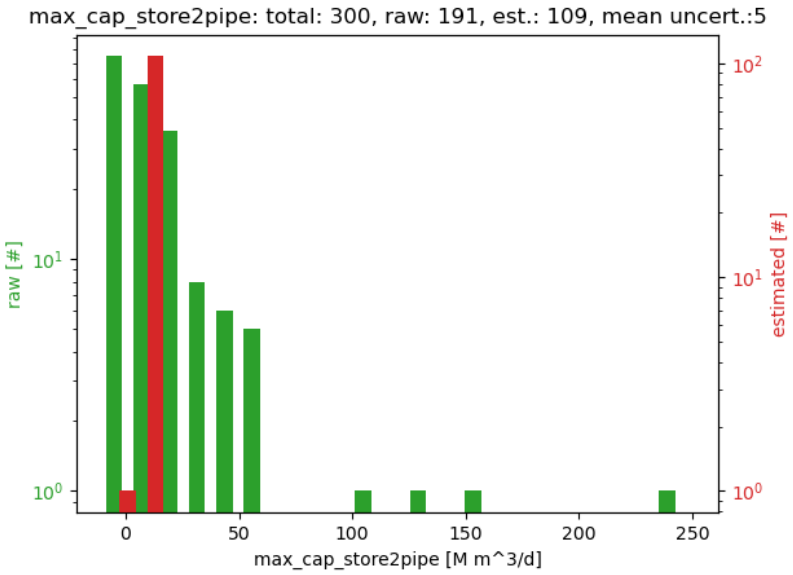


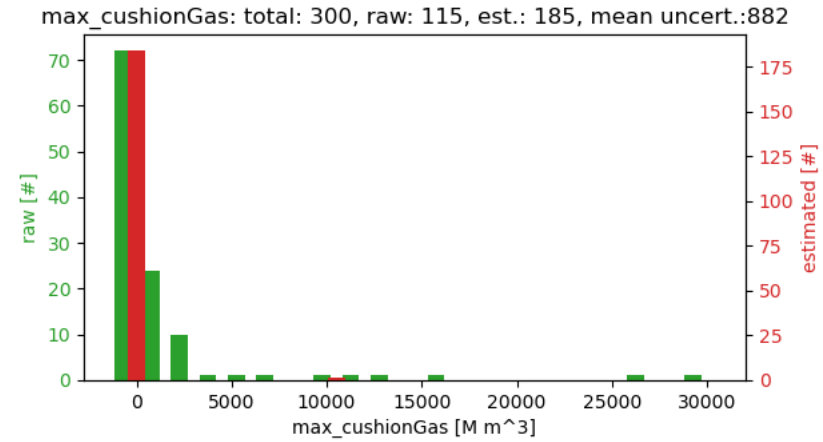
Median(max_cap_pipe2store_M_m3_per_d): 118 (mean uncert.: 11)
Lasso(max_cap_store2pipe_M_m3_per_d, Pipe_max_cap_M_m3_per_d): 1 (mean uncert.: 3)
min(raw):1.09, max(raw):244.84, min(est):2.85, max(est):7.28



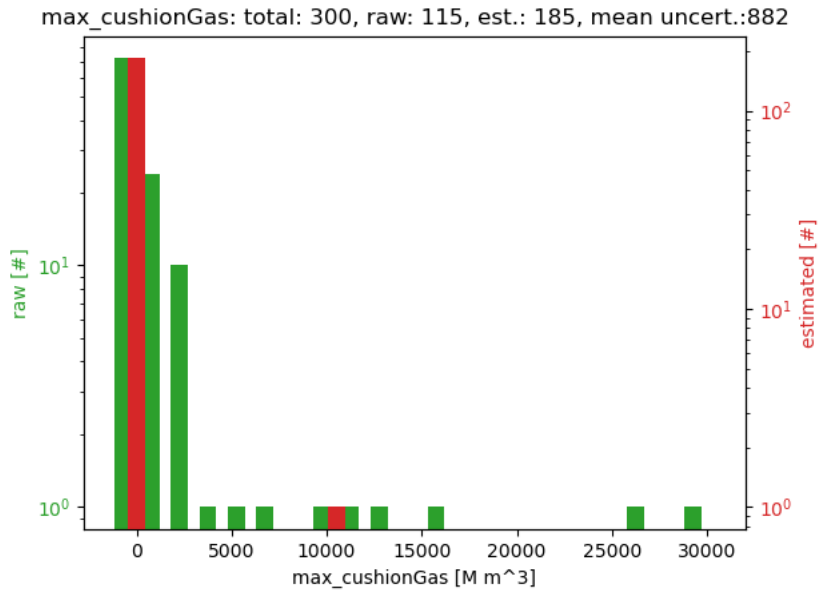


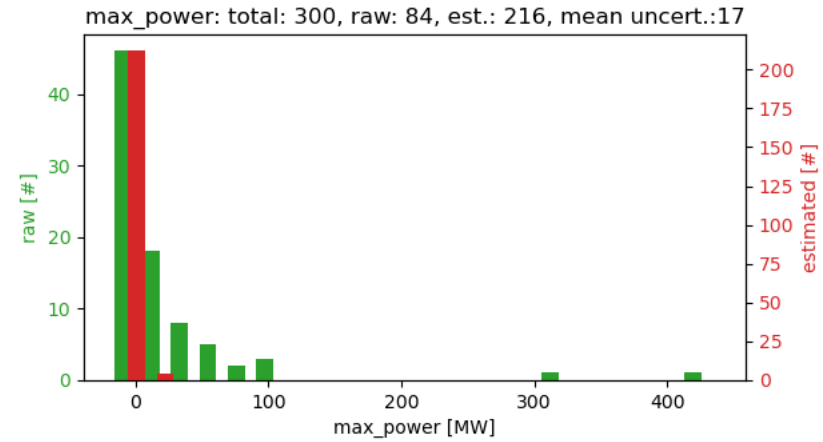
Median(max_cap_store2pipe_M_m3_per_d): 104 (mean uncert.: 13)
Lasso(max_cap_pipe2store_M_m3_per_d, max_workingGas_M_m3): 1 (mean uncert.: 6)
Lasso(max_workingGas_M_m3, Pipe_max_pressure_bar): 4 (mean uncert.: 6)
min(raw):1.01, max(raw):244.84, min(est):4.36, max(est):10.84



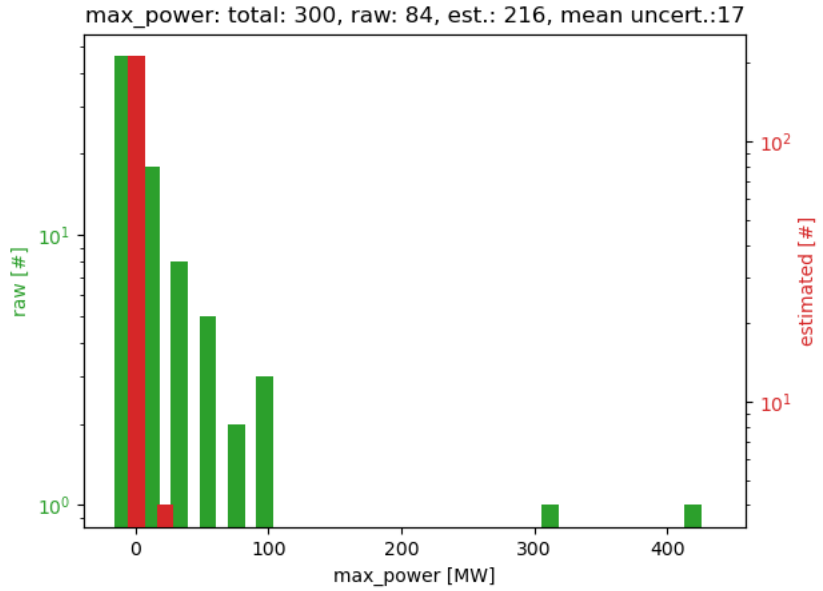


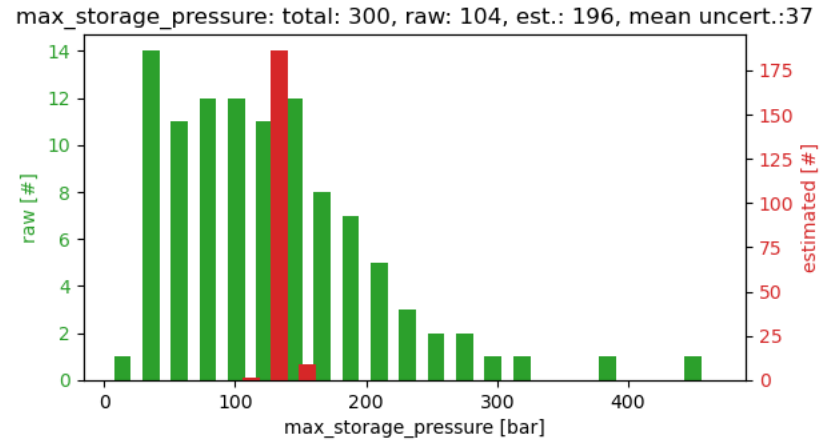
Median(max_cushionGas_M_m3): 164 (mean uncert.: 1595)
Lasso(num_storage_wells, Pipe_max_pressure_bar): 21 (mean uncert.: 134)
min(raw):1.8, max(raw):30000.0, min(est):92.29, max(est):10274.9



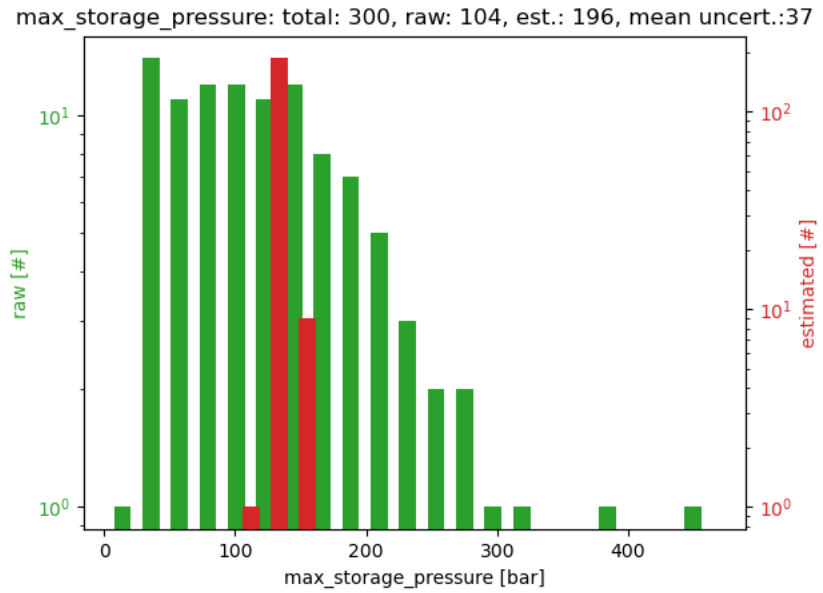


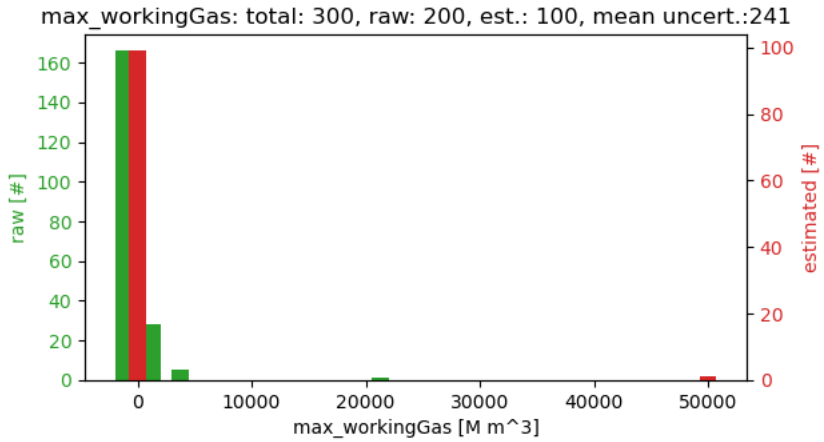
Median(max_power_MW): 198 (mean uncert.: 24)
Lasso(num_storage_wells, Pipe_max_cap_M_m3_per_d): 18 (mean uncert.: 6)
min(raw):1.0, max(raw):430.0, min(est):8.24, max(est):19.96



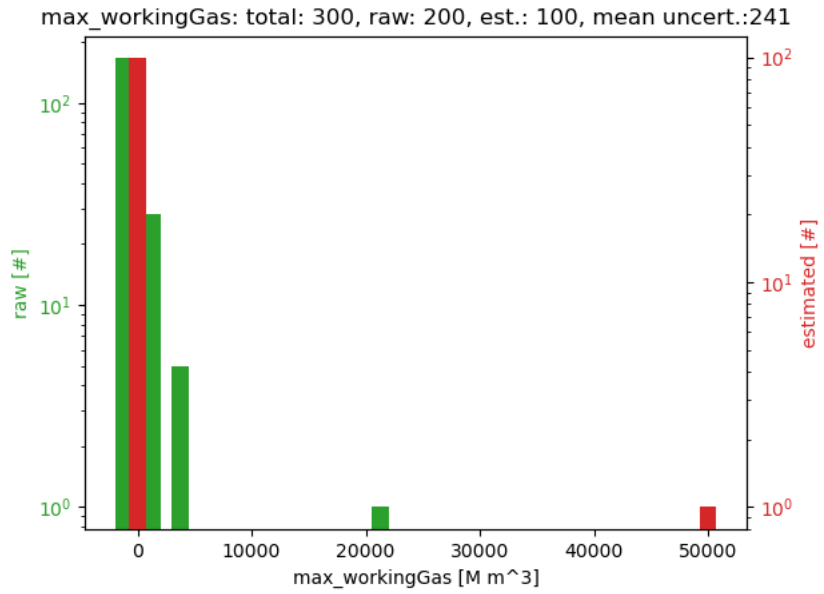


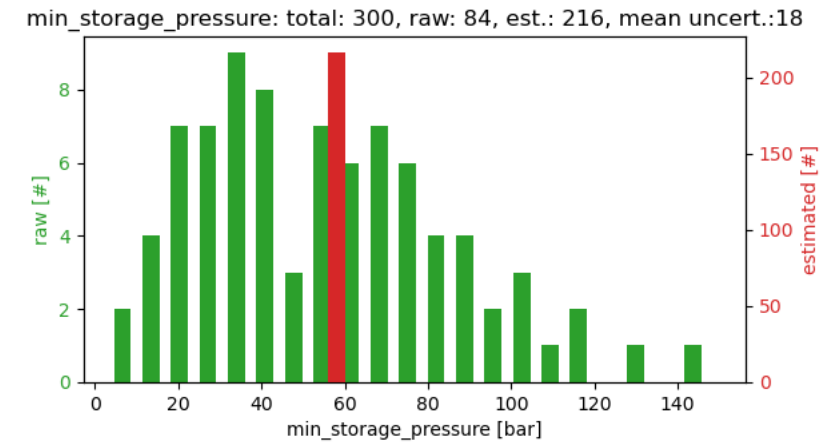
Median(max_storage_pressure_bar): 171 (mean uncert.: 59)
Lasso(min_storage_pressure_bar, Pipe_diameter_mm): 25 (mean uncert.: 39)
min(raw):25.0, max(raw):460.0, min(est):121.94, max(est):155.63



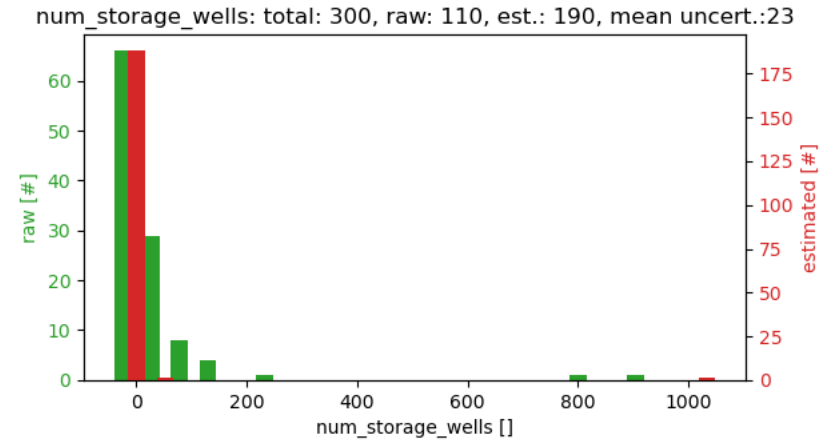


Median(max_workingGas_M_m3): 95 (mean uncert.: 649)
Squared(num_storage_wells, Pipe_max_pressure_bar): 4 (mean uncert.: 121)
M_Heuristics.adjust_AttribValues(): 1 (mean uncert.: 10000)
min(raw):11.4, max(raw):23000.0, min(est):170.19, max(est):50000.0

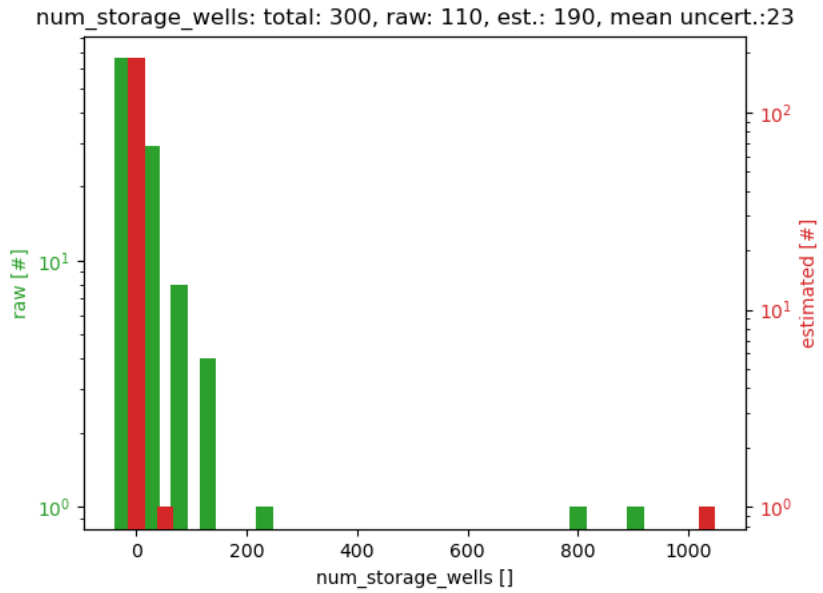




Median(min_storage_pressure_bar): 216 (mean uncert.: 25)
min(raw):10.0, max(raw):147.0, min(est):60.0, max(est):60.0



Median(num_storage_wells): 172 (mean uncert.: 40)
Lasso(max_workingGas_M_m3, Pipe_max_pressure_bar): 17 (mean uncert.: 6)
Squared(max_cap_store2pipe_M_m3_per_d, Pipe_max_pressure_bar): 1 (mean uncert.: 7)
min(raw):1, max(raw):941, min(est):9.84, max(est):1033.88



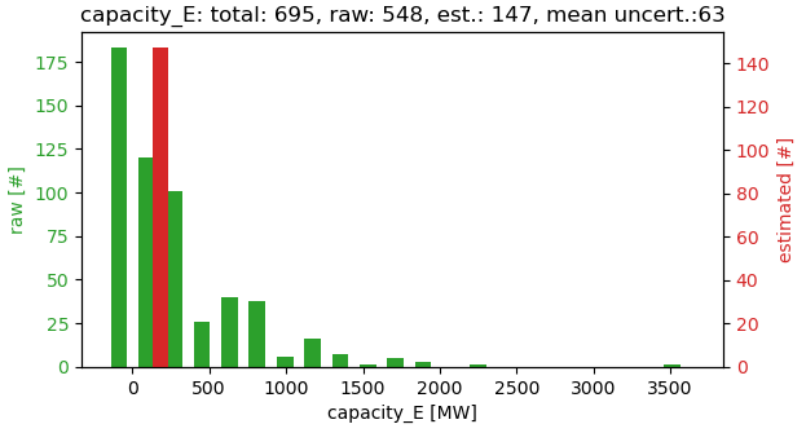
4.6.6 *Consumers*

As there was no data generated through the heuristic processes for the elements of *Consumers*, there is no need to compare attribute distributions.

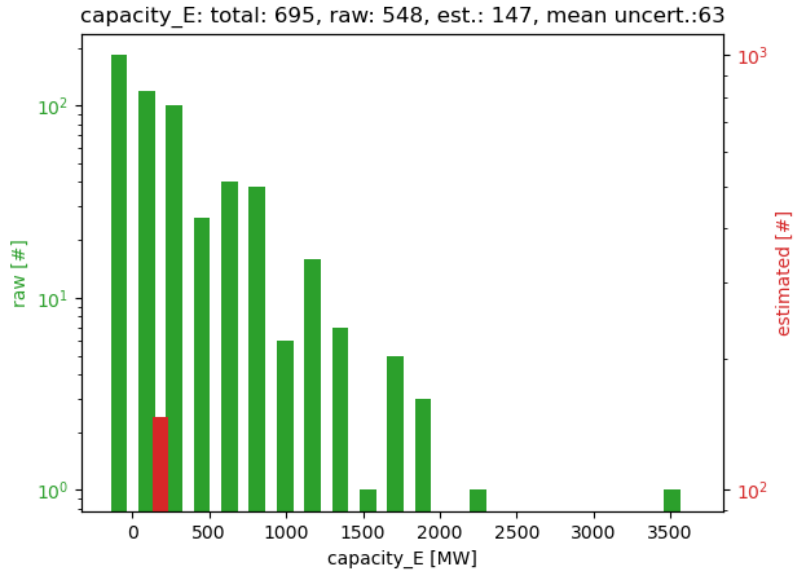
4.6.7 *PowerPlants*

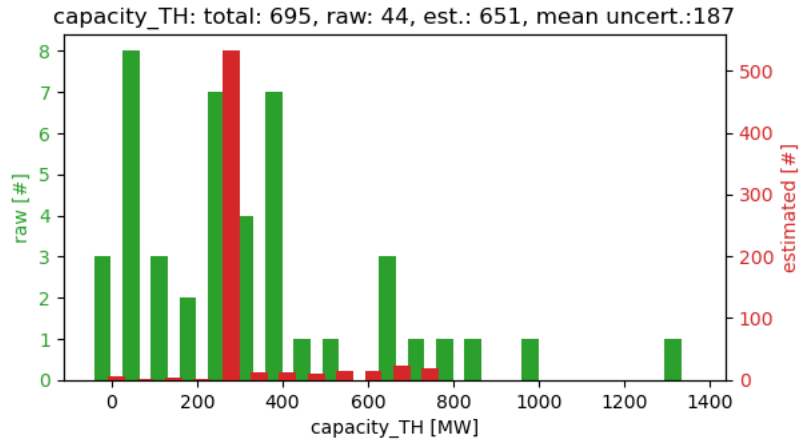
Below are the heuristic histogram plots of the component *PowerPlants* for the attributes:

- *capacity_E_MW*
- *capacity_TH_MW*

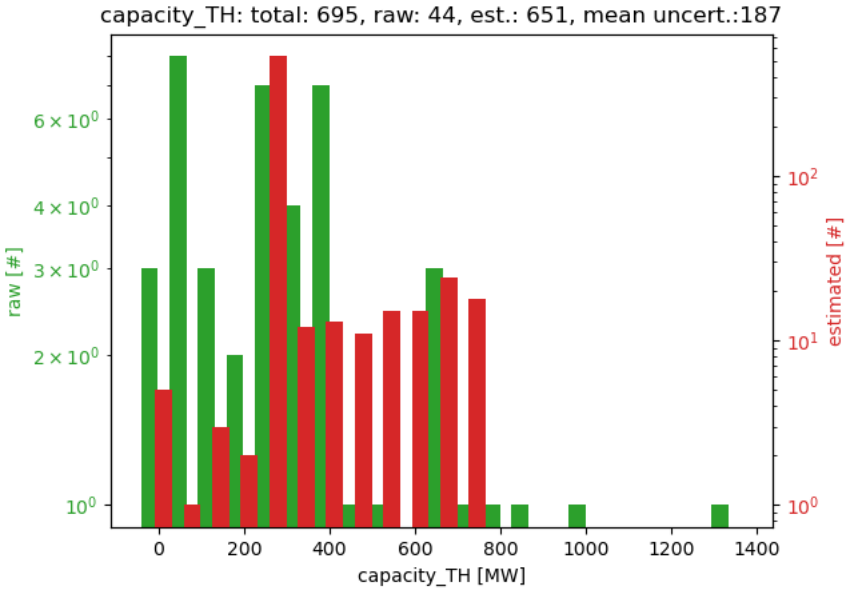


Median(capacity_E_MW): 147 (mean uncert.: 297)
min(raw):2.9, max(raw):3600.0, min(est):199.0, max(est):199.0





Median(capacity_TH_MW): 507 (mean uncert.: 214)
Lasso(Pipe_diameter_mm, Pipe_max_cap_M_m3_per_d): 139 (mean uncert.: 117)
M_Heuristics.adjust_AttribValues(): 5 (mean uncert.: 1000)
min(raw):25.0, max(raw):1347.0, min(est):12.0, max(est):759.77



4.7 Acknowledgement

We acknowledge the contribution of Dr. Ontje Luensdorf from the German Aerospace Center (DLR), Institute for Networked Energy Systems to the SciGRID_gas project.

BIBLIOGRAPHY

[AFW14] M. Ahmed, B.T. Fasy, and C. Wenk. *New Techniques in Road Network Comparison*. Penguin Random House, New York, NY, 2014.

[AG99] H. Alt and L. Guibas. *Discrete geometric shapes: matching, interpolation, and approximation-a survey*. Sack JR, Urrutia J, Handbook of Computational Geometry, Elsevier, New York, NY, 1999.

[CCCS21] 14 06 2018. [Online] Copernicus Climate Change Service. Era5 hourly data on single levels from 1979 to present. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview, 2021. Accessed: 2021-01-15.

[Die21] J.C. Diettrich. Generation of a non-osm scigrid_gas gas transmission network data set. gitHub, 2021. to be published as part of the code release.

[DPDi20] J.C. Diettrich, A. Pluta, J. Dasenbrock, and W. i. *SciGRID_gas: The combined IGGIN gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, 2020. URL: https://zenodo.org/record/4288459#.YG7aFj9CSUk, doi:10.5281/zenodo.4288458.

[DPi20] J.C. Diettrich, A. Pluta, and W. i. *SciGRID_gas: The INET gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, Aug 2020. URL: https://doi.org/10.5281/zenodo.4008975, doi:10.5281/zenodo.4008975.

[DPM20a] J.C. Diettrich, A. Pluta, and W. Medjroubi. *SciGRID_gas: The combined IGG gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, Aug 2020. URL: https://doi.org/10.5281/zenodo.4009129, doi:10.5281/zenodo.4009129.

[DPM20b] J.C. Diettrich, A. Pluta, and W. Medjroubi. *SciGRID_gas: The combined IGGI gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, Aug 2020. URL: https://zenodo.org/record/4288468#.YG7aBz9CSUk, doi:10.5281/zenodo.4288467.

[DPM20c] J.C. Diettrich, A. Pluta, and W. Medjroubi. *SciGRID_gas: The combined IGGIELGN gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, 2020. URL: https://zenodo.org/record/4642569#.YG7ZhT9CSUk, doi:10.5281/zenodo.4642568.

[DPM20d] J.C. Diettrich, A. Pluta, and W. Medjroubi. *SciGRID_gas: The combined IGGINL gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, 2020. URL: https://zenodo.org/record/4288440#.YG7aHz9CSUk, doi:10.5281/zenodo.4288439.

[DPM20e] J.C. Diettrich, A. Pluta, and W. Medjroubi. *SciGRID_gas: The raw INET data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, Aug 2020. URL: https://doi.org/10.5281/zenodo.3985249, doi:10.5281/zenodo.3985249.

[DPM20f] J.C. Diettrich, A. Pluta, and W. Medjroubi. *SciGRID_gas: The raw LKD data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, Aug 2020. URL: https://doi.org/10.5281/zenodo.3985271, doi:10.5281/zenodo.3985271.

[DPS+21] J.C. Diettrich, A. Pluta, J.E. Sandoval, J. Dasenbrock, and W. Medjroubi. *SciGRID_gas: The combined IGGIELGNC-3 gas transmission network data set*. DLR-Institut für Vernetzte Energiesysteme e.V., Oldenburg, Germany, 2021. URL: https://zenodo.org/record/4922530#.YNLX7kxCS60, doi:10.5281/zenodo.4922529.

[Ent17] EntsoG. North West GRIP, Min Report. https://www.entsog.eu/sites/default/files/files-old-website/publications/GRIPs/2017/entsog_GRIP_NW_2017_main_xs.pdf, 2017. Accessed: 2021-03-17.

[Eur21a] EuroStat. Average size of dwelling by household type and degree of urbanisation. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_hcmh02&lang=en, 2021. Accessed: 2021-02-08.

[Eur21b] EuroStat. Complete energy balances. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_bal_c&lang=en, 2021. Accessed: 2021-02-08.

[Eur21c] EuroStat. Disaggregated final energy consumption in households - quantities. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_d_hhq&lang=en, 2021. Accessed: 2021-02-07.

[Eur21d] EuroStat. Employment by age, economic activity and nuts 2 regions (nace rev. 2). https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lfst_r_lfe2en2&lang=en, 2021. Accessed: 2021-02-08.

[Eur21e] EuroStat. Gross domestic product (gdp) at current market prices by nuts 3 regions. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nama_10r_3gdp&lang=en, 2021. Accessed: 2021-03-31.

[Eur21f] EuroStat. Number of households by degree of urbanisation and nuts 2 regions. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lfst_r_lfsd2hh&lang=en, 2021. Accessed: 2021-02-08.

[Eur21g] EuroStat. Population on 1 january by age group, sex and nuts 3 region. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_r_pjangrp3&lang=en, 2021. Accessed: 2021-01-10.

[Eur21h] EuroStat. Sbs data by nuts 2 regions and nace rev. 2 (from 2008 onwards). https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=sbs_r_nuts06_r2&lang=en, 2021. Accessed: 2021-03-16.

[Gas19] GasTerra. Aardgas in Nederland. https://www.gasterra.nl/uploads/fckconnector/20d5b75a-2ad6-529e-b024-1f5bd43bbd5e, 2019. Accessed: 2021-03-17.

[GGB+20] F. Gotzens, B. Gillessen, S. Burges, W. Hennings, J. Müller-Kirchenbauer, S Seim, P. Verwiebe, T. Schmid, F. Jetter, and T. Limme. *Harmonization and development of methods for a spatial and temporal resolution of energy demands (DemandRegio)*. Forschungsstelle für Energiewirtschaft e.V., 2020. Accessed: 2021-05-18.

[Hel18] D. Helle. OpenStreetMap - Deutschland. https://www.openstreetmap.de/, 2018. Accessed: 2019-12-12.

[Kha13] Y. Khalid. What is Pickle in python? https://pythontips.com/2013/08/02/what-is-pickle-in-python/, 2013. Accessed: 2019-10-10.

[KingSpalding18] King&Spalding. LNG in Europe 2018: An Overview of LNG Import Terminals in Europe. https://globallnghub.com/wp-content/uploads/2018/09/King.pdf, 2018. Accessed: 2018-09-01.

[KKS+17] F. Kunz, M. Kendziorski, W.-P. Schill, J. Weibezahn, J. Zepter, C. von Hirschhausen, and P. Hauser. *Electricity, Heat, and Gas Sector Data for Modeling the German System*. Deutsches Institut für Wirtschaftsforschung, Daten Dokumentation 92, Berlin, 2017.

[LSS+19] P. Lustenberger, F. Schumacher, M. Spada, P. Burgherr, and B. Stojadinovic. Assessing the performance of the european natural gas network for selected supply disruption scenarios using open-source information. *Energies*, 12(4685):1–28, 2019. doi:{10.3390/en12244685}.

[MMK16] C. Matke, W. Medjroubi, and D. Kleinhans. SciGRID - An Open Source Reference Model for the European Transmission Network (v0.2). https://power.scigrid.de, 2016. Accessed: 2019-09-09.

[Mic20] Microsoft. Bing maps api. https://www.microsoft.com/en-us/maps/licensing, 2020. Accessed: 2018 to 2021.

[Nis20] A. Nisbet. Open topo data api. https://www.opentopodata.org, 2020. Accessed: 2020 to 2021.

[San21] Javier Enrique Sandoval. Estimation and simulation of a gas demand time series for the european nuts 3 regions. Master's thesis, Carl von Ossietzky Universität Oldenburg, Germany, Fak. 5, Institute of Physics (PPRE) D-26111 Oldenburg, Germany, 6 2021. Supervised by Prof. Dr. Carsten Agert, Dr. Herena Torio and Dr. Wided Medjroubi.

[San19] B. Sandvik. World Borders. http://thematicmapping.org/downloads/world_borders.php, 2019. Accessed: 2019-07-07.

[SAB+17] M. Schmidt, D. Aßmann, R. Burlacu, J. Humpola, I. Joormann, N. Kanelakis, T. Koch, D. Oucherif, M.E. Pfetsch, L. Schewe, R. Schwarz, and M. Sirvent. *GasLib—A Library of Gas Network Instances*. 2017. doi:{10.3390/data2040040}.

[sl19] scikit-learn. 1.1. Linear Models (scikit learn). https://scikit-learn.org/stable/modules/linear_model.html, 2019. Accessed: 2019-08-08.

[UoO14] USA University of Oregon. Comparing distributions: Z Test. http://homework.uoregon.edu/pub/class/es202/ztest.html, 2014. Accessed: 2020-07-07.

[Wik20a] Wikipedia. Bootstrapping (statistics). https://en.wikipedia.org/wiki/Bootstrapping_(statistics), 2020. Accessed: 2019-06-06.

[Wik20b] Wikipedia. Cross-validation (statistics). https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Exhaustive_cross-validation, 2020. Accessed: 2019-07-07.

[Wik20c] Wikipedia. Jackknife resampling. https://en.wikipedia.org/wiki/Jackknife_resampling, 2020. Accessed: 2019-08-08.

[Wik20d] Wikipedia. Lasso (statistics). https://en.wikipedia.org/wiki/Lasso_(statistics), 2020. Accessed: 2020-04-04.

[Wik20e] Wikipedia. Limited-memory BFGS. https://en.wikipedia.org/wiki/Limited-memory_BFGS, 2020. Accessed: 2020-06-06.

[Wik20f] Wikipedia. Out-of-bag error. https://en.wikipedia.org/wiki/Out-of-bag_error, 2020. Accessed: 2019-07-07.

[Wik20g] Wikipedia. Transmission system operator. https://en.wikipedia.org/wiki/Transmission_system_operator/, 2020. Accessed: 2019-09-09.

[Wik20h] Wikipedia. JAGAL. https://en.wikipedia.org/wiki/JAGAL, 2020. Accessed: 2020-01-01.

[Wik21] Wikipedia. Nomenclature of territorial units for statistics. https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics, 2021. Accessed: 2021-04-10.

[BMWi11] BMWi. Forschung für eine umweltschonende, zuverlässige und bezahlbare Energieversorgung. https://www.bmwi.de/Redaktion/DE/Publikationen/Energie/6-energieforschungsprogramm-der-bundesregierung.pdf?__blob=publicationFile&v=12, 2011. Accessed: 2019-02-02.

[BMWi20] BMWi. Home page of BMWi. https://www.bmwi.de/Navigation/DE/Home/home.html, 2020. Accessed: 2020-03-03.

[BundesregierungDeutschland20] Bundesregierung Deutschland. Home page of Bundesregierung Deutschland. https://www.bundesregierung.de/Webs/Breg/DE/Themen/Energiewende/_node.html, 2020. Accessed: 2020-01-01.

[EntsoG20] EntsoG. Home page of EntsoG. https://www.entsog.eu/, 2020. Accessed: 2020-03-03.

[GasIEurope20] Gas Infrastructure Europe. Home page of Gas Infrastructure Europe. https://agsi.gie.eu, 2020. Accessed: 2020-01-01.

[GasSEurope20] Gas Storages Europe. Home page of Gas Storages Europe. https://www.gie.eu/index.php/transparency/gse-transparency-template, 2020. Accessed: 2020-01-01.

[Gassco20a] Gassco. Data page of facilities from Gassco. https://www.npd.no/en/about-us/information-services/available-data/map-services/, 2020. Accessed: 2020-01-01.

[Gassco20b] Gassco. Home page of Gassco Norway. https://www.gassco.no/en/, 2020. Accessed: 2020-01-01.

[IGU20] IGU. Home page of International Gas Union. https://www.igu.org/, 2020. Accessed: 2018-10-01.

[nationalGrid20] nationalGrid. Home page of National Grid UK. https://www.nationalgrid.com/uk/, 2020. Accessed: 2018-10-01.